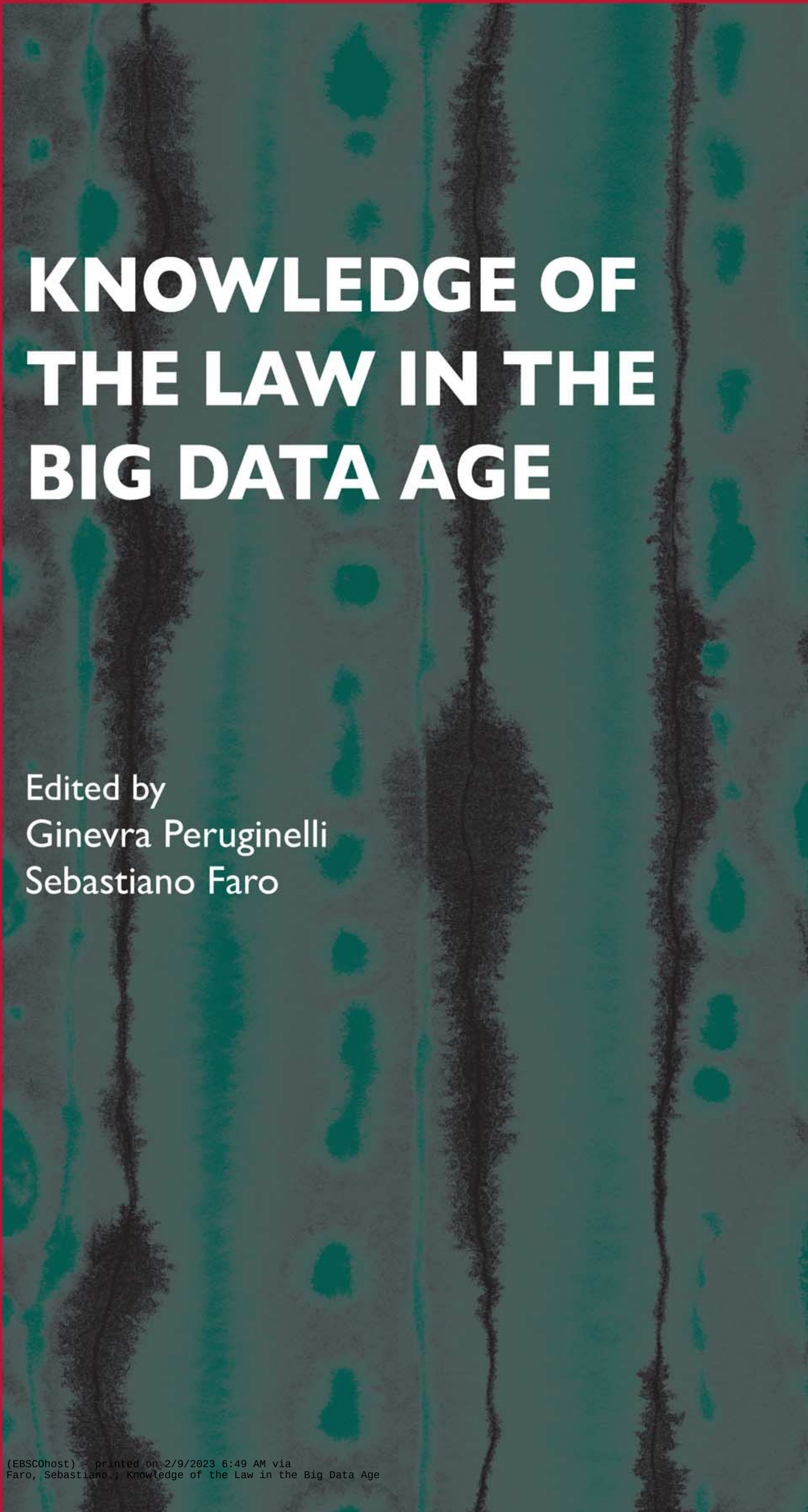
Frontiers
in
Artificial
Intelligence
and
Applications

KNOWLEDGE OF THE LAW IN THE BIG DATA AGE

Edited by
Ginevra Peruginelli
Sebastiano Faro



IOS
Press

KNOWLEDGE OF THE LAW IN THE BIG DATA AGE

The changes brought about by digital technology and the consequent explosion of information known as Big Data have brought opportunities and challenges in all areas of society, and the law is no exception.

This book, Knowledge of the Law in the Big Data Age contains a selection of the papers presented at the conference 'Law via the Internet 2018', held in Florence, Italy, on 11-12 October 2018. This annual conference of the 'Free Access to Law Movement' (http://www.fatlm.org) hosted more than 60 international speakers from universities, government and research bodies as well as EU institutions.

Topics covered range from free access to law and Big Data and data analytics in the legal domain, to policy issues concerning access, publishing and the dissemination of legal information, tools to support democratic participation and opportunities for digital democracy. The book is divided into 3 sections: Part I provides an introductory background, covering aspects such as the evolution of legal science and models for representing the law; Part II addresses the present and future of access to law and to various legal information sources; and Part III covers updates in projects, initiatives, and concrete achievements in the field.

The book provides an overview of the practical implementation of legal information systems and the tools to manage this special kind of information, as well as some of the critical issues which must be faced, and will be of interest to all those working at the intersection of law and technology.

9 781614 999843

ISBN 978-1-61499-984-3 (print)
ISBN 978-1-61499-985-0 (online)
ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

KNOWLEDGE OF THE LAW IN THE BIG DATA AGE

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Volume 317

Recently published in this series

Vol. 316. S. Borgo, R. Ferrario, C. Masolo and L. Vieu (Eds.), Ontology Makes Sense – Essays in honor of Nicola Guarino
Vol. 315. A. Lupeikiene, O. Vasilecas and G. Dzemyda (Eds.), Databases and Information Systems X – Selected Papers from the Thirteenth International Baltic Conference, DB&IS 2018
Vol. 314. L.C. Jain, X. Zhao, V.E. Balas and F. Shi (Eds.), Information Technology and Intelligent Transportation Systems – Proceedings of the 3rd International Conference on Information Technology and Intelligent Transportation Systems (ITITS 2018) – Xi'an, China, September 15–16, 2018
Vol. 313. M. Palmirani (Ed.), Legal Knowledge and Information Systems – JURIX 2018: The Thirty-first Annual Conference
Vol. 312. T. Endrjukaite, A. Dudko, H. Jaakkola, B. Thalheim, Y. Kiyoki and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XXX
Vol. 311. M. Coeckelbergh, J. Loh, M. Funk, J. Seibt and M. Nørskov (Eds.), Envisioning Robots in Society – Power, Politics, and Public Space – Proceedings of Robophilosophy 2018 / TRANSOR 2018
Vol. 310. N. Petkov, N. Strisciuglio and C. Travieso-González (Eds.), Applications of Intelligent Systems – Proceedings of the 1st International APPIS Conference 2018

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Knowledge of the Law in the Big Data Age

Edited by

Ginevra Peruginelli

Institute of Legal Information Theory and Techniques
National Research Council of Italy (ITTIG-CNR)

and

Sebastiano Faro

Institute of Legal Information Theory and Techniques
National Research Council of Italy (ITTIG-CNR)

IOS
Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-61499-984-3 (print)
ISBN 978-1-61499-985-0 (online)
doi: 10.3233/FAIA317

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

LEGAL NOTICE

The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Preface

Free online access to information is approaching maturity and is evolving in line with the Big Data ecosystem: data volumes are continuing to grow and so are the possibilities of what can be done with so much raw data available. The major challenges of the Big Data age are of course well known (volume of ever-increasing data, variety of data types and structures, contribution of big data to evidence-based decision making). Added to these is not only the increasing number of disciplines and problem domains where Big Data is having effects, but also the consequent challenges and opportunities for Big Data to have a major impact on science, business, and government. In recent times, the legal domain and in particular legal information management has started to embrace this trend for better accessing, disseminating and understanding law, for improved decision making, and so much more.

Big Data represents both the greatest innovation and peril of these times and promises to provide a scientific and empirical approach to law. The digital paradigm has revolutionized the ability to communicate legal information: the physical and geographical inhibitors no longer matter. Legal information must be available on the Internet, should be freely available, and nobody should have to pay to access information essential to one's rights and obligations in a working democracy. A specific issue also involves distribution or redistribution of legal information and the way that it is accomplished. In this regard not only are there obligations on the part of the governments, but upon those who are responsible for the distribution of and access to legal information.

The topics addressed in this volume are situated in such a context, and range from free access to law, Big Data, data analytics in the legal domain, to policy issues for accessing, publishing and disseminating legal information, as well as tools to support democratic participation and opportunities for digital democracy.

These aspects, together with other related issues, have been specifically tackled and discussed at the international conference 'Law via the Internet 2018' which was organized by the Institute of Legal Information Theory and Techniques of the National Research Council of Italy (CNR-ITTIG) on 11–12 October 2018 in Florence. This was the annual conference of the 'Free Access to Law Movement' (http://www.fatlm.org) which brings together over 60 Legal Information Institutes (LIIs) from all over the world. The Florentine Conference hosted more than 60 speakers from universities, government and research bodies as well as EU institutions, who animated a lively and wide debate on the main theme that gave the specific title to the Conference and gives also now the title to this book: 'Knowledge of the Law in the Big Data Age'.

The volume collects a selection of papers presented in Florence by Italian and foreign experts who accepted our invitation to contribute. The structure of the book reflects partially the sections designed for the Conference. Part I, entitled 'Encountering Big Data and Law', provides an introductory background, covering foundational issues such as legal epistemology, future of machine-driven evolution of legal science and (semi)formal models for representing the law.

Part II, dedicated to 'Challenges and Opportunities in Disseminating and Accessing Legal Information', addresses the present and future of access to law, with particular reference to ‘Rules, Policies and Publication Models’ for disseminating the increasingly rich and varied legal information sources and to the analysis of ‘Standards and interoperability’ issues.

Part III, ‘Experiences, Good Practices and Critical Issues’, is about essential updates in projects, initiatives, and concrete achievements in the field. The result is a picture that, far from being exhaustive, provides an overview of practical implementation of legal information systems, tools to manage this special kind of information and some critical issues to face.

The volume arises from the idea to reflect on the actual methods and strategies to access and have knowledge of the law as it is today and to compare the ways through which it is distributed and is made accessible. The intent is to present the current state of the discussion and to offer new perspectives of reflection on issues central in the current debate on the relationship between law and technology. As editors, our hope is, according to our research interests, that, by going through the chapters of this volume, the reader may realize how successful initiatives in the direction of free access to law mainly depend on at least two factors: the strong potential of interdisciplinary collaboration in today’s web environment, and the capacity of legal culture to understand and meet the challenges of the Big Data age.

We want to express many thanks to all the contributors who make this volume precious well above the merits of the editors. They deserve our gratitude for having patiently waited for the completion of this collective work. Moreover, we are glad to publish the volume with open access: this is in line with the desire for protecting free access to legal knowledge and confirms our firm belief and strong support of the philosophy of open access.

Finally, be informed that from June 1st, 2019, the Institute of Legal Information Theory and Techniques (CNR-ITTIG) which we, as researchers, are part of, will change its name to Institute of Legal Informatics and Judicial Systems (CNR-IGSG). We are particularly happy to have been able to go out with this last precious volume that leads us to a new and exciting chapter of our Institute’s life, full of new expectations, stimuli and challenges.

Florence, 30 May 2019 *Ginevra Peruginelli* and *Sebastiano Faro*

Acknowledgement: We would like to express our special thanks to Giuseppina Sabato and Simona Binazzi from CNR-ITTIG for their precious editing work of this volume.

Contents

Section II.2. Interoperability and Standards

Part III. Experiences, Good Practices and Critical Issues

Part I

Encountering Big Data and Law

This page intentionally left blank

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190003

Legal Epistemology in the Times of Big Data

Vincenzo ZENO-ZENCOVICH
Università degli Studi Roma Tre (Italy)

Abstract. This Chapter presents some of the main challanges put to lawyers by the growing Big Data environment. In particular it points out what are the consequences of passing from a causal logic to an inferential logic.

Keywords. big data, datasphere, visual analytics, philosophy of law

1. Introduction

Is there a '*Beruf unserer Zeit*' in the times in which Big Data, Internet-of-Things, and Data Driven Innovation appear to be magic words that capture the attention of scientists, economists, sociologists and obviously also of jurist?

Clearly – and by now thousands of pages demonstrate it – the main effort has been that of pouring the new digital wine in the age-old wineskins, mostly those of private law.

But there appears to be a much more engaging task that requires to be, at least, pointed out, leaving to the slow meditation of legal doctrine its more deep analysis and framing.

The question is if this relatively new technological scenario, which is growing at an impressive speed and which apparently cannot be stopped, changes the traditional epistemological bearing of lawyers, i.e the way they understand the world and offer legal tools for its best functioning.

2. Size Matters

Data have always existed and public institutions and private entities have been collecting and using them for millennia. This phenomenon obviously increases with the development of modern societies and the use of data-capturing technologies. Where is the significant change?

The main problem is that the sheer size of the data, that can be measured only in trillions of trillions and ever growing, requires new tools to understand them, extract knowledge, use them. These tools not only go beyond traditional epistemology, but tend to suggest predictively what might happen. The law has been for centuries mostly deontic. Now it becomes, increasingly an instrument to put into place forecasts that are envisaged through data analytics.

The second most significant change is in what may be called the ‘T factor’. In the past, policy – and therefore normative – choices were taken on the basis of past data, statistics, a term whose roots indicate a ‘static’ situation. Nowadays, especially with the development of Internet-of-Things technologies data are collected, processed and analyzed in real-time.

This implies that decisions – also legal decisions – be taken in real time on the basis of the available data (typically, closing a route because of an accident or congestion; embargo of import of a product for sanitary reasons). In order to be timely, legal procedures – mostly of an administrative nature – will have to pre-ordered on the basis of predictive programmes, which operate automatically according to a simplified logic. Whatever procedural cautions may be considered necessary, they will be applied in a prior moment, that of the elaboration of the programme. Once this has been implemented the legal consequences will automatically ensue.

As often happens with scientific and technological issues, the level of understanding that lawyers – especially those placed in decision-making institutions, Parliament, administrations, courts – have is rather limited. In part because the problems are complex and not easily unraveled. In part because there is an innate tendency of lawyers of deference towards technical decisions.

But is this still possible? If data become the basis of automated decisions of a legal significance (in the first place decisions by public authorities, but also of private parties) [1]; [2] it appears necessary to understand their nature, the logics that are behind their analysis, the undisclosed presumptions and biases [3].

What needs to be stressed is that data analytics is intrinsically value-ridden. Data, in themselves and especially if one is talking of billions of data, do not speak out, but in general provide a certain, and variable, answer according to the question one is putting.

Hypothetically the same data can be used to detect consumption propensity of a community, for epidemiological purposes, or to organize public transport needs.

In order to understand if the use of data – and of data analytics – is correct one has to investigate, and question, the quality of data i.e. their nature, the methods through which they have been collected, the algorithms used to elaborate them.

It is necessary to point out that in social practices one is still lacking a consensus on the appropriate methodologies and even less on codes of conduct[1].

Such a situation seems to suggest the need for some kind of hard or soft regulation. At any rate, a deferential attitude appears to be of no use.

Finally, on the same line, one should discuss, evaluate and decide if the analytical conclusions drawn from certain data can be applied also to other social spheres and public decisions. Looking at rather common regulations (e.g. building permits; environmental precautions; commercial activities) lawyers are well aware of the differences and of the fact that rules in one field are, logically and practically, inapplicable in others. This awareness needs to be applied to the world of data and to their regulatory use[2] [6].

[1] For an attempt see the document voted on 12 January 2017 by the American Association for Computing Machinery, *Statement on Algorithmic Transparency and Accountability*, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf); and [4]; [5].

[2] According to [7]: “Advances in technology (such as big data and artificial intelligence) will give rise to this new form – the micro-directive – which will provide the benefits of both rules and standards without the costs of either. Lawmakers will be able to use predictive and communication technologies to enact complex legislative goals that are translated by machines into a vast catalog of simple commands for all possible scenarios”.

3. The Creation of a 'Datasphere'

One has pointed out that the size of the data is not simply a technical issue that can be solved through digital resources. Surely one needs more advanced computing resources in order to collect and store the ever increasing number of data. The issue, however, appears to be more complex [8].

The datafication of real life objects has until now been one of the most common ways through which data is created. A printed book is scanned, the digital photograph of an object is taken, the sound of an event (whether a conversation, a musical event, a natural phenomenon) is recorded, etc. Roughly speaking, one can imagine that through this process the whole material world is, or can be, datafied, creating a parallel digital world. A good example of such duplicity is the Google Street View service which offers a reproduction, from many different perspectives, of a town. Although in some cases it is intended to work as a map, the amount of information contained in the data goes well beyond simple directions and virtually transports the viewer to that specific spot.

The digitalisation of objects, although far from being completed, concerns a duplication of the real world, a duplicate that can be used anytime and anywhere by an unlimited number of people, adapted to one's needs or interests, and transformed in something new and different[3].

Seeing things from outside, this means that objects in the material world, once datafied, have what might be considered a 'second life', and the two 'spaces' constantly refer to one another. Additionally, although digital versions of material objects (e.g. of a book, of a painting, of a musical performance) have existed for decades, putting them in a broader environment of interaction with other data illustrates their complexities, and also their significant differences.

This new entity one may call 'Datasphere' and is the digital replica of the world and of what is known of the universe has its own rules. One is not advocating the libertarian view of the Internet as a utopian space subtracted from the interference of national and supra-national laws. One has learnt, by now, that cyberspace is an intensively regulated environment, in its infrastructure, in its modes of communication, in its content.

What one wishes to point out is that the datasphere, which one could metaphorically describe as an ocean in which continuously telecommunication networks pour, every instant, millions of data, extract other millions in a never-ending flow, is very scarcely governable.

The significant difference is that while oceans are relatively stable, the data sphere is ever-growing at an accelerate speed.

Not only the data-producing apparatuses are increasing in the million (from energy meters to automobiles; from agricultural control sensors to domotic appliances), but data, one poured into the datasphere produce further data in a never ending process.

One could say that while humans are attempting to discover the universe, defying its infinity, at the same they are creating, on Earth, a potentially infinite digital environment made of data.

What are the implications for the law?

The first obvious remark is that governing this digital world is not the same as governing the natural world.

[3] One can find sufficient elements in the vast report by the OECD, *Data-driven Innovation for Growth and Well-being* (October 2014), and see how much way has been made over the last five years.

Setting rules for human conduct is different from setting rules (which?) that should control mostly automated processes in a digital environment.

The second remark is related to the ubiquity or the undetectable location of the data. It is extremely difficult to establish where the data is held and if it is all held in only one space. The issue has clear implications as to the applicable law which is not only – or even in prevalence – private law but more often public law, in its multiple administrative and regulatory aspects. In fact, while there is an ample body of law and of jurisprudence dealing with conflicts of laws and of jurisdictions in private law matters, there is no similar solution for administrative law which inevitably can apply only within the boundaries of the national jurisdiction and is unable to be effective outside it [9].

For the sea there are specific international conventions. For the datasphere one has still to establish what it is.

Finally, one must consider that the datasphere appears to be governed – differently from natural elements like the sea, airspace or outer space – entirely by technological means (hardware, software, networks for connection etc.) created by humans. This, in a certain way, could suggest that once one establishes what humans can, or cannot, do in the datasphere, an orderly system of rules could be put in place. However, not only are these rules of an entirely technical nature, but what is more significant is that the whole system is intrinsically based on self-learning procedures. Although we are still far from replicating the complexity of a human mind, yet artificial intelligence, and therefore the ability of machines to replicate mental processes through completely autonomous acquisition of data is difficult to stop or regulate.

Lawyers are not afraid of novel worlds – whether geographical or artificial – and have the tools to tackle them and put order in what appears to be chaos.

This however requires an appropriate use of taxonomies, avoiding the pitfalls of digital jargon or of juvenile slang. The first challenge is therefore of identifying situations, taking into account the rapid changes in technology. While it is clear that first come the phenomena, and then comes the law, one cannot avoid reminding non-lawyers that logic and formal language are not the privilege of computer science but also fundamental columns of the building of the law. Just as in the former field one refuses fuzzy notions that do not bring anywhere, also in the latter one must define, distinguish, order.

The law could be qualified as a proto-software and this is the reason why legal informatics have developed so rapidly as a specific branch of knowledge and of research.

On the other hand, one must avoid using existing legal institutions (property, contract, administrative procedures) as a Procrustes' bed on which to lie, forcibly, the elements of the new technological world.

Other concerns, however, appear to be more compelling. The law is a social product, and it cannot exist if it is not followed, voluntarily or through effective enforcement, by a community. Digital technologies not only allow the law to be known and applied anytime, anywhere, but also suggest, and even impose, such kind of use.

One can therefore envisage a growing move towards automated systems of the application of the law, already much in use in the sectors of taxation and of social security, both in the relations between individuals and authorities, and between individuals. The combination of legal rules with data bring into our perspective so-called 'granular norms' or, to use an expression which actually reveals its opposite, 'smart contracts'.

Norms lose their general scope and are 'tailored' on the identity of the individual – mostly revealed by the personal data one dispones about the person. One could and should investigate the role that predictive analytics have in the enforcement of the law. The most noticeable example is their use which is already widely in place, in police investigations in order to narrow the groups of suspected persons, or establish links between suspects and unsuspected individuals. On a much more every-day level, predictive analytics are used – and increasingly will be – in the selection of employees, in insurance contracts, in school and university admission tests, and in other contexts in which individual rights and a transparent relationship with public authorities is at stake.

4. Epistemological Concerns

One has already pointed out that Big Data and their environment have already changed the way the law is and is implemented. A few further points should be highlighted [10]; [11].

- For millennia legal reasoning has been based on what may be called causal logic, which is well epitomized by the couple IF/THEN[4].

  Big Data, because of their volume, of their variety, of the velocity at which they are collected and elaborated are generally searched through inferential logic, expressed in the couple IF/MAYBE. This clearly opens the road to multiple solutions, but generally only the one which is most probable, on the basis of the available data, is considered.

  While this does not present particular concerns in the digital world (the best example is that of online shopping portals: 'Customers who have bought A have also considered buying B'), when applied to legal norms we find it challenges our deeply rooted principles. One does not apply a rule to everybody simply because it is the most probable or majoritarian situation. Each rule is composed of several elements more or less strictly defined (very strictly for criminal law, less strictly in other cases). If one or more of those are lacking, the rule will not apply. To take into account the inevitable differences in the various cases there is a clear relationship between general rule/general exceptions/special rule.

  Inferences sweep all this away and suggest a different way of thinking and acting. Will it become common in legal thinking to?
- Big Data for their sheer size are difficult to understand and the results of searches on them difficult to convey. This has brought to the development of visual analytics or infographics. These colorful depictions catch the eye and are supposed to convey the sense of the data. This raises several doubts. In the first place one wonders if such presentations actually do represent the results of the research on the data, and if for sake of neatness and of aesthetics the whole picture has been over-simplified. In the second place, communicating through visual instruments is

[4] See Plato's *Meno*: "A statue that is tied down, though, is very valuable, because the man's works are very beautiful. What am I driving at here? True opinions. True opinions, for as long as they remain, are fine things and do nothing but good. But they don't hang around for long; they escape from a man's mind, so that they are not worth much until one tethers them with chains of causal reasons. And these, Meno my friend, are threads of memory, as previously agreed. After opinions are tied down, in the first place they become knowledge; secondly, they remain in place. That is why knowledge is prized more highly than correct opinion; knowledge differs from correct opinion in being tied down".

different from trying to explain, through words, to the mind[5]. If one applies the ideal/visual dichotomy to legal reasoning and rules, one must doubt that the law – and its complexities – can be expressed through pictures. And that the endless variety of case law – which is the first source of legal Big Data – can correctly and faithfully represented through data visualization [12].

- The final issue raised by the Big Data revolution is the emergence of new class of rule-makers. Until the mid-20th century, policy decisions were mostly taken by individuals who had a legal education and knew 'how to do things with rules'. Then economists entered the scene, profoundly influencing the turn of great or small policies. Nowadays, this role is increasingly taken by computer scientists whose professional status is unclear, whose role is rarely transparent, and whose activity does not appear to be subject to professional and deontological standards. In other terms, a norm is known by the public and anybody can challenge it on the basis of other norms. However, this is rarely the case if behind the norm there are data and algorithms. Data accountability becomes therefore a central issue in any modern democracy: but this accountability is not only of a technical nature but also, and primarily, of a deontic nature [13]. What are the biases that govern the inferences in data analytics: age, sex, race, location etc.? What are the assumptions upon which predictive analytics are grounded? What, if any, are the goals pursued: Efficiency? Fairness? Equality? Social justice?

References

[1] Supiot, A. (2015). *La gouvernance par les nombres. Cours au Collège de France 2012–2014*. Librairie Fayard.

[2] Hansen, H. K. & Porter, T. (2017). What Do Big Data Do in Global Governance? *Global Governance, 23*, 31.

[3] Bass, G. D. (2015). Big Data and Government Accountability: An Agenda for the Future. *I/S: A Journal of Law and Policy for the Information Society, 11*(1), 13.

[4] Rosenblat, A., Kneese, T. & Boyd, D. (2014). *Algorithmic Accountability*, https://datasociety.net/pubs/2014-0317/AlgorithmicAccountabilityPrimer.pdf.

[5] Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review, 165*, 633.

[6] Mitts, J. (2014). *Predictive Regulation*, https://ssrn.com/ abstract=2411816.

[7] Casey, A. J. & Niblett, A. (2017). The Death of Rules and Standards. *Indiana Law Journal, 92*, 1401.

[8] Bergé, J-S., Grumbach, S. & Zeno-Zencovich, V. (2018). The 'Datasphere', Data Flows beyond Control, and the Challenges for Law and Governance. *European Journal of Comparative Law and Governance, 5*(2), 144-178.

[9] de Jong-Chen, J. (2015). Data Sovereignty, Cybersecurity, and Challenges for Globalization. *Georgetown Journal of International Affairs, 16*, 112.

[5] Aristotle, *On the soul*, in parts 3, 7, 8 and 9 analyses in detail the relation between thought, understanding, knowledge, images and imagination, to the point that "the soul never thinks without an image". And in *On interpretation* (part 1) "Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images". And D. Hume, in his *Enquiry Concerning Human Understanding* (1748): "Perceptible objects always have a greater influence on the imagination that anything else does, and they readily convey this influence to the ideas to which they are related and which they resemble" (Section 5); and "Philosophy teaches us that images (or perceptions) are the only things that can ever be present to the mind, and that the senses serve only to bring these images before the mind and cannot put our minds into any immediate relation with external objects" (Section 12).

[10] Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society, 1*(1), 1-12, http://eprints.maynoothuniversity.ie/5364/1/RK_big%20data.pdf.

[11] Frické, M. (2015). Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology, 66*(4), 651-661.

[12] Zeno-Zencovich, V. (2018). Through a Lawyer's Eyes. Data Visualization and Legal Epistemology. In Degrave, E., de Terwangne, C., Dusollier, S. & Queck, R. (Eds.). *Law, Norms and Freedoms in Cyberspace*. Larcier, 459.

[13] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190004

Knowledge Machineries. Introducing the Instrument-Enabled Future of Legal Research and Practice

Nicola LETTIERI
Istituto Nazionale per l'Analisi delle Politiche Pubbliche - INAPP; Università degli Studi del Sannio - Dipartimento di Diritto, Economia, Management e Metodi Quantitativi (Italy)

Abstract. 30 years after the advent of the World Wide Web, information and communication technologies keep triggering deep changes in the way we access, produce and use knowledge. The convergence between data warehouse facilities and computational science heuristics is populating the Internet with cloud infrastructures designed to manage and process information in completely new ways. We are facing the emergence of a new generation of online platforms integrating knowledge management, data analytics, visualization and collaboration tools for purposes that gradually move from information retrieval to scientific research. This Chapter introduces the looming of the platform era in the legal world showing, also by means of concrete examples, how these tools can be used to make the most of the growing amount of legal information today accessible online. The analysis becomes an opportunity to dwell on how computational tools can turn into the emergence of new perspectives in legal research and practice.

Keywords. legal analytics, machine science, digital platforms

1. Introduction

In the era of planetary-scale computation, ICT are offering much more than the direct access to huge amounts of information. The gradual integration of data sharing facilities, online collaboration, computational heuristics, and visualization is bringing about the emergence of online infrastructures providing a wide range of innovative services that go far beyond information retrieval. Digital platforms are not only shaping sociality, economies, politics, and institutions [1] but also transforming the very way knowledge is produced, organized and shared in any field of human activity. This is real whether we think about the access to endless repositories of documents, or scientific research.

Even if in peculiar ways, this process involves law as well. Scholars and practitioners are getting to grips with the creation of platforms[1] as they appear to be a promis-

[1] A case in point – mainly focused on lawyers' perspective – is represented by a talk published online by the *LawLab* (an interdisciplinary teaching and research center focused on legal innovation and technology of the Illinois Tech - Chicago-Kent College of Law): Kennedy, D. (2017, December). *Agile Lawyering in the Platform Era* [Video file], https://vimeo.com/246985325.

ing way to enhance the knowledge of the law in the Big data era. The challenges are many: besides the still existing need to fine-tune technical standards allowing to manage legal documents, we have before us the opportunity of exploiting innovative heuristics to extract better knowledge from the increasing amount of legal information today available. A reference point, in this perspective, lies in the emerging data and computation-driven research paradigm in which every stage of the scientific endeavour, from questions definition to the sharing of results, is enhanced by digital information-processing, computational heuristics and distributed collaboration infrastructures [2]. Technological and methodological findings from the area of ‘machine science’ [3] can not only inspire new solutions for legal information retrieval or access to law but also open new scenarios in terms of research questions.

The Chapter sets out a critical introduction to the opportunities of a machine-driven evolution of law and legal research. The analysis, that brings together technical, methodological and epistemological issues, is structured as follows. Section 2 provides a brief overview of the emerging ‘machine science’ paradigm focusing on the role played by online research platforms as tools for data analysis and theory development. Section 3 and 4 introduce the appearance of online analytical platforms in legal field also by referring to some ongoing research projects. Section 5 focuses on a research project aiming to exploit machine learning and visualization to gain new knowledge from both legal information and administrative microdata. The last Section will sketch some considerations about the role of computational and data-driven tools in the evolution of legal science and practice. In the background, the idea that, together with other cultural and methodological factors, they can trigger much-needed changes in the scope, aims and methods of legal research: an empirical and interdisciplinary turn, and an opening towards a methodologically eclectic approach to scientific investigation.

2. Beyond Information Retrieval: Online Platforms and Machine Science

Science is facing a machine-driven future. Cloud computing, open/big data, artificial intelligence and a growing legion of computational heuristics, are pushing researchers toward an ever more symbiotic relationship with machines. Last 20 years have witnessed the development of a series of approaches to scientific research that, regardless of the name adopted – ‘computational science’ [2]; [3]; [4], ‘data science’ [5]; [6], ‘e-science’ [7]; [8] and, more recently, ‘machine science’ [3] – can be all traced back to a vision attaching to data and, above all, computational tools a prominent role for our understanding of reality.

In this perspective, technologies work as ‘epistemic enhancers’ [2]: they extend our natural observational and cognitive abilities reframing research practices and the very way we deal with information and knowledge [9]; [10]; [11]. Changes brought by the machine-driven approach to science are significant. If a first fundamental one has been the spread of computer simulations in new subject areas [12]; [13], a second profound change has been the data-driven turn of scientific inquiry. Indeed, once coupled with the power of computation, the volume and variety of information today available have triggered the rise of a further research paradigm [5]; [10] whereby scientific hypoteses are preceded – when not completely replaced – by the identification of hidden patterns in huge amounts of data, rather than stemming from observation made accordingly to an explicit theoretical model [11].

In a time of such profound changes, computational methodologies are spreading not only in empirical research – mainly through Big Data analytics and machine learning – but also in theory-making – mostly by means of computer simulation model building. Scientist's toolkit is steadily expanding, and research has a growing need for new tools seamlessly integrating the building blocks of the data and computation-driven scientific paradigm. In this scenario, digital platforms – infrastructures consisting of hardware, software, networking systems and data management components to perform computationally demanding tasks – are becoming the *sine qua non* of innovative practices in which technological infrastructures are used to support potentially all the stages of the research path from questions definition to interactive data exploration and visualization, experiment modeling, data analysis and sharing of the findings (Figure 1).

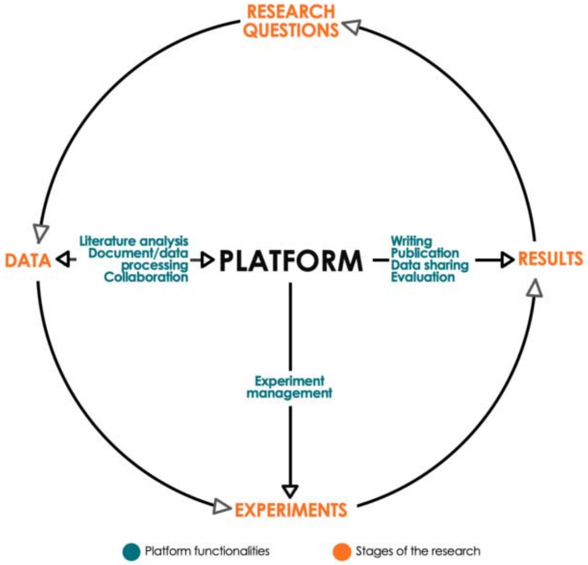


Figure 1. The machine-science research cycle

Actually, the machine-driven evolution of research is already a fact. Last ten years have been marked by the spread of a number of platforms aimed at supporting research in many different ways. The taxonomy proposed in [14], gives an idea of the variety of solutions emerged so far at the same time offering food for thoughts for further applications (Table 1).

Along with the computational science perspective, digital platforms are set to have a disruptive impact on social science and humanities that, for inherent features and historical reasons, are less familiar with quantitative and instrumental research practices than natural science. While growing amounts of information about substantially all social phenomena are stored in digital archives, they are pushing social sciences towards a paradigm in which the mainly qualitative analyses are increasingly backed up with data-driven and computational solutions. This way, the work of any social scientist is going to

Table 1. A taxonomy of digital platforms for science

Category	Description
Literature Analysis	*Goal*: *i*) help researchers in exploring the growing amount of scientific papers available online with ad hoc search engines; *ii*) ease the navigation within the materials integrating visualisation, bookmarking and publication-sharing system. *Examples: Bibsonomy, CiteUlike, Google Scholar, Mendeley, ReadCube, PubChase.*
Data and Code Sharing	*Goal*: *i*) support the management of large sets of data and programming code allowing to collect, share, cite and reuse materials in social and natural science. *Examples: Github, CodeOcean; Socialsci, GenBank; DelveHealth, BioLINCC.*
Collaboration	*Goal*: *i*) facilitate researchers in reaching out to other researchers and find expertise for scientific cooperation; *ii*) communicate research activities to the general public; *iii*) involve the general public in the research efforts according to the 'citizen science' paradigm (e.g. by sharing CPU time or classifying pictures). *Examples: Academia, ResearchGate, Loop; Kudos* and *AcaWiki.*
Experiments	*Goal*: *i*) help researchers in all the activities connected with scientific experiments: equipment and data management; scheduling; research protocols; coding and data analysis; generating and analysing data; visualising results. *Examples: Asana, LabGuru; Quartzy, Transcriptic, GitLab, Wolfram Alpha, Sweave*
Writing	*Goal*: *i*) support paper drafting helping researchers to write papers with other people while keeping track of modifications made by authors on the document; *ii*) allowing researchers to share with colleagues bibliographies, citations, and references. *Examples: Endnote, Zotero, Authorea, ShareLaTex, Citavi.*
Publication	*Goal*: *i*) ease the publication and discussion of papers accelerating scientific interactions and discovery; *ii*) allow authors to connect papers with additional materials like executable code. *Examples: GigaScience, Cureus, ArXiv, Exec&Share, RunMyCode.*
Research Evaluation	*Goal*: *i*) enhance research evaluation supporting in new ways paper review and analysis of the impact of publications. *Examples: PubPeer, Publons, Academic Karma, Altmetric, ImpactStory.*

include also the effort to imagine and experiment with innovative tools, 'places' where theories, data, computational heuristics can converge, be explored[2].

Basically the same applies to the legal world. To give just an example, as suggested by recent developments in Computational legal studies, the integration of legal data and sophisticated processing pipelines is becoming crucial to derive relevant knowledge from legal documents. Projects using machine learning to predict the behaviour of supreme courts [15] or exploiting computational heuristics to assess the structural and semantic complexity of US legal corpora [16] witness the interest in innovative solutions to analyze legal information. In front of this, it makes sense to state that a significant part of future efforts in legal research will have to be devoted to the creation of tools allowing to extract actionable knowledge from data. What is at stake goes further than the effi-

[2] As highlighted by Kitchin [11], the emerging field of computational social science provides an important opportunity to develop more sophisticated models of social life allowing scientists to shift "from data-scarce to data-rich studies of societies from static to dynamic unfoldings; from coarse aggregations to high resolutions; from relatively simple models to more complex, sophisticated simulations".

ciency in carrying out traditional research practices but involves, actually, the potential emergence of new ways to delve into the complexity of legal world [17].

3. The Rise of Legal Machineries: A Brief Background

The development of integrated tools for the legal research has already been a reality for some time. There are several examples of fully-fledged systems – some of which online – dealing with the analysis of legal data and documents. On the other hand, the need for new 'analytical machineries' is looming on today's debate about aims and methods of legal studies. As a matter of fact, two flourishing research areas, Empirical legal studies (ELS) [18]; [19]; [20] and Computational legal science [21]; [22]; [23], are somehow pushing forward debate on the instruments by which law can be explored and studied. The call for a closer integration of empirical analyses into legal scholarship that characterises ELS, for example, inevitably results into the quest for data-driven tools and practices enhancing our understanding of law as a fact [24]. Likewise, even if with different research goals, computational legal scholars are working hard to figure out new ways to make the most of digital tools and heuristics. Current scenario shows different experiences heading in this direction, some linked in a specific way with the study of legal texts, others more oriented to empirical analysis. Here below, in a non-exhaustive manner, a brief overview of some of these experiences giving a sense of the ongoing trends in the development of those we define as 'legal machineries'.

LexMex[3] is a simple but interesting online system exploiting visualization techniques to represent the relations between texts of law for purposes of information retrieval and study. The attention is focused on the French Civil Code and related legislation. The tool generates a graph transforming laws in nodes and citations in edges. The semantics of the visualization is simple: the size of node depends on the number of connections it has with other nodes. Colours correspond to the cluster detected by means of community-detection algorithm allowing to identify groups of highly connected norms. The tool implements essential navigation such as zooming, node selection to show contextual information, and search by keywords.

Ravel Law[4] is a legal analytics research tool, a new type of search engine that combines analytics (natural language processing, machine learning), legal research and graph visualization to help finding the way through a comprehensive case law database from the Harvard Law Library. Unlike traditional legal databases, presenting search results in textual form, using long lists that often hide important cases pages back in search results, *Ravel Law* visually depicts the most important cases on a particular topic as nodes of a network, with edges pointing to subsequent cases citing it. The size of the nodes is used to represent the relevance of the precedents estimated using proprietary algorithms. Besides bringing about a change in how legal research is presented online, *Ravel Law* has also contributed to an interesting and still ongoing debate about the use of 'next gen' research tools in legal education [25].

The use of large collections of digitized texts and citation network analysis to come up with new insightfull methodologies for legal studies is discussed in [26]. Authors have developed an open source software for the analysis and visualization of networks

[3] http://www.lexmex.fr/.

[4] https://home.ravellaw.com/.

of Dutch case law, aiming to support both legal scholars and non-technical researches in their investigations. The basic goal of the research is to answer in new ways legal research questions, including those of determining relevance of case law precedents, comparing the precedents with those identified in the literature, and determining clusters of related cases. In a similar direction, again, it is possible to cite a work presented in 2016 exploiting visual approaches to depict and explore the history of Swiss Federal Law [27]. Authors wonder whether the degree of norms complexity can be measured over time and how it can be represented. To answer this question, they have organized in a same pipeline OCR, parsing (to obtain structured XML from textual documents), data analysis (in particular complexity measures like *Shannon entropy of word use; depth of the hierarchical structure* and *density of external references*), and visualizations.

Turning to the more 'empirically oriented' tools mentioned above, we could cite a number of works. Many experimental software systems have been developed, for example, that combine empirical data from criminal investigations (wiretaps, online communications, environmental tappings), data mining and visualization to support criminal court judges, public prosecutors and law enforcement agencies in the fight against crime [28]; [29]; [30]; [31]. Due to space limitations, we confine ourselves to cite just one recent and illustrative example of this category of tools. *SIIMCO* [32] is a forensic investigation software suite for identifying the influential members of a criminal organization. The system exploits data provided by public prosecutors and police departments (in particular, crime incident reports and mobile communication data about the members of the organization) to create network diagrams representing the structure of criminal organizations. *SIIMCO* employs then network analysis measures to quantify the degree of influence/importance of each individual, to detect subgroups, discover interaction patterns between groups, and identify central members.

4. Fiddling with Legal Analytical Platforms

To make concrete our speculations about the role of analytical platforms in enhancing legal information, we briefly present in the following the rationale and results of an ongoing experimental activity carried out along with the Department of Computer Science of the University of Salerno[5]. The research, still ongoing, aims to gain a first-hand experience with the prospects opened up, in the legal world, by the platform paradigm. Over the last four years, the initiative has already turned into the development of a series of experimental online environments in various ways dealing with legal data, knowledge mining and visualization. The choice to develop from scratch *ad hoc* tools despite the availability of many existing solutions (see, for instance, the variety of research platforms listed in Table 1), makes sense if we turn our mind to a series of needs arising both in legal research and practice:

- *Customization*: tailored solutions (algorithms, workflows) can better fit both the nature (structure, characteristics, errors etc.) of data handled and the research goals.

[5] The projects have seen the involvement of people from law (scholars, lawyers, public prosecutors), computer science, visualization, computational biology.

- *Openness*: from scratch development allows to avoid proprietary software increasing intelligibility and comparability of algorithms and easing the analysis and sharing of the results.
- *Integration*: custom-designed tools make it easier to integrate in one place different functionalities (e.g different kinds of visualizations) and heuristics (e.g. network analysis, machine learning, agent-based modeling etc.).

Based on the consideration of these needs, we have been working on the development of a series of platforms that variously integrate legal data, computational heuristics and visualization to support activities spanning from legal knowledge mining to scholarly research. In more details, experiments headed in the following research directions: enhance legal information retrieval; extend and step up the methodological apparatus available to legal scholars interested in the empirical analyses; figure out new ways to identify and measure the computational correlates of legal concepts (e.g. relevance of case law precedents). Here below a brief description of the tool realized over the years.

- *KnowLex* [33] is a web application designed to allow a more intuitive exploration and analysis of documents coming from different legal sources. The goal of the tool is to exploit visual analytics techniques to support the understanding of the legal framework on a given issue when, as often happens, this requires the analysis of complex legal corpora.
- *EUCaseNet* [34] is an online analytical platform in various ways allowing legal scholars to explore the features of the entire body of European Court of Justice case law: i) by applying network analysis metrics (centrality measures, *Page Rank*, community detection algorithms) to its citation network so to study, for instance, the relevance of precedents; ii) by exploiting statistical visual analytics tools applied to case law metadata.
- *CrimeMiner* [35] is an experimental knowledge mining platform exploiting empirical and legal data from real criminal proceedings (crime incident reports, wiretaps, environmental tappings, criminal records etc.) to support legal practitioners (public prosecutors, judges, law enforcement agencies) and scholars (computational social scientists, criminologists) in investigating the features of criminal networks and of their members. To this end the tool integrates into an ever evolving pipeline information extraction, graph analytics, agent-based social simulation [36] and, in a recent experiment, machine learning.

In more recent times, drawing on the wealth of technical experiences and methodological findings stemming from the projects just above mentioned, we have started a new initiative exploiting legal information and administrative microdata to support both the access to public sector information and data-driven policy design. A more detailed description is offered in the following section.

5. Argos: Visualization and Machine-learning with Normative and Administrative Data

Due also to the spread of the open government paradigm [37], recent years have been marked by a growing effort of public administrations to start the extensive collection and sharing of data generated within their institutional activities. Large-scale administrative data today show high levels of quality in terms of temporal resolution, volume, and structure [38]. Their diffusion and exploitation raise a growing deal of attention for different reasons. According to EU strategies for the development of the information society[6], they must be made easily available not only to increase the transparency of government and administrative activities but also because, when integrated with other public sector data and analysed with appropriate heuristics, they form the backbone of any evidence-based policy making and agenda setting. In this scenario, it is still difficult fully exploit the potential of data also due to the lack of tool offering tylored analytics and advanced interaction-design solutions in open source format. Challenges for researchers wishing to take advantage of large dataset are different: gaining access to data, developing data management and programming capabilities needed to work with large-scale datasets, and finally thinking of creative approaches to summarize, describe, and analyze information.

Argos is a project that points in this direction by developing a modular online platform allowing the visualization and analysis of large amounts of administrative, legal and economic data. The goal of the project is twofold: i) facilitate the interaction with large-scale administrative data using infographics that make access, comprehension and re-elaboration of information by experts and citizens easier and more intuitive; ii) experiment machine-learning techniques to extract actionable knowledge from cross-cutting reading of heterogeneous (administrative and normative) data. Among the target users there are for sure legal scholars, especially those interested in supporting their studies (e.g. those needed for regulatory impact analysis) with reliable estimates about correlations between regulatory measures and social/economic impacts. The platform prototype includes two modules, both still under development, that are being tested using administrative microdata relating to the impact of the labour market reforms adopted in Italy from 2008 up to 2015[7].

(1) Visualization module

The module (already online, see Figure 2) provides two visualizations that allow to interactively explore the dataset offering insights about the evolutions in the structure of the Italian labour market. More in details, the visualizations can be described as follows: i) the *Zoomable treemap* offers an intuitive and interactive navigable representation of the proportions that bind the elements of groups selected and hierarchically ordered by the user so to answer questions like: How many of the employment contracts signed from 2008 to 2015 are open-ended (OE) or fixed-term (FT)? To what extent education levels are associated with each of the two types of work relationships? ii) the *Bubble Chart/GIS map* allows to intuitively explore the evolution of multidimensional phenomena in order to make evident the fluctuations, over time, of variables that are supposed to be somehow correlated. Our attention was focused on representing trends in OE and FT employment contracts in the period 2008-2015, at the same time offering other information useful to interpret the phenomenon both in an economic and in a legal perspective. To this end, we put in the same visualization a set of different information: trends of OE/FT contracts in individual Italian regions; trend of OE/FT contracts in the North/Center/South areas; enactment of the main labour market reforms; start and end dates of the legislatures. The

[6] The Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information represents the starting point a regulatory process that has developed through different regulatory measures. Among the others, it is worth mentioning the Communication of the EU Commission *Open Data. An Engine for Innovation, Growth and Transparent Governance* COM (2011) 882.

[7] Year of the so called ‘JOBS Act’, a set of regulatory measures that, among the other things, have introduced in the Italian legal system permanent hiring subsidy and new regulations lowering firing costs.

Bubble Chart is complemented with a GIS map offering a geo-referenced and animated representation of data on labour market trends, so to allow users to visually relate the normative, spatial and temporal dimensions of the changes underway.

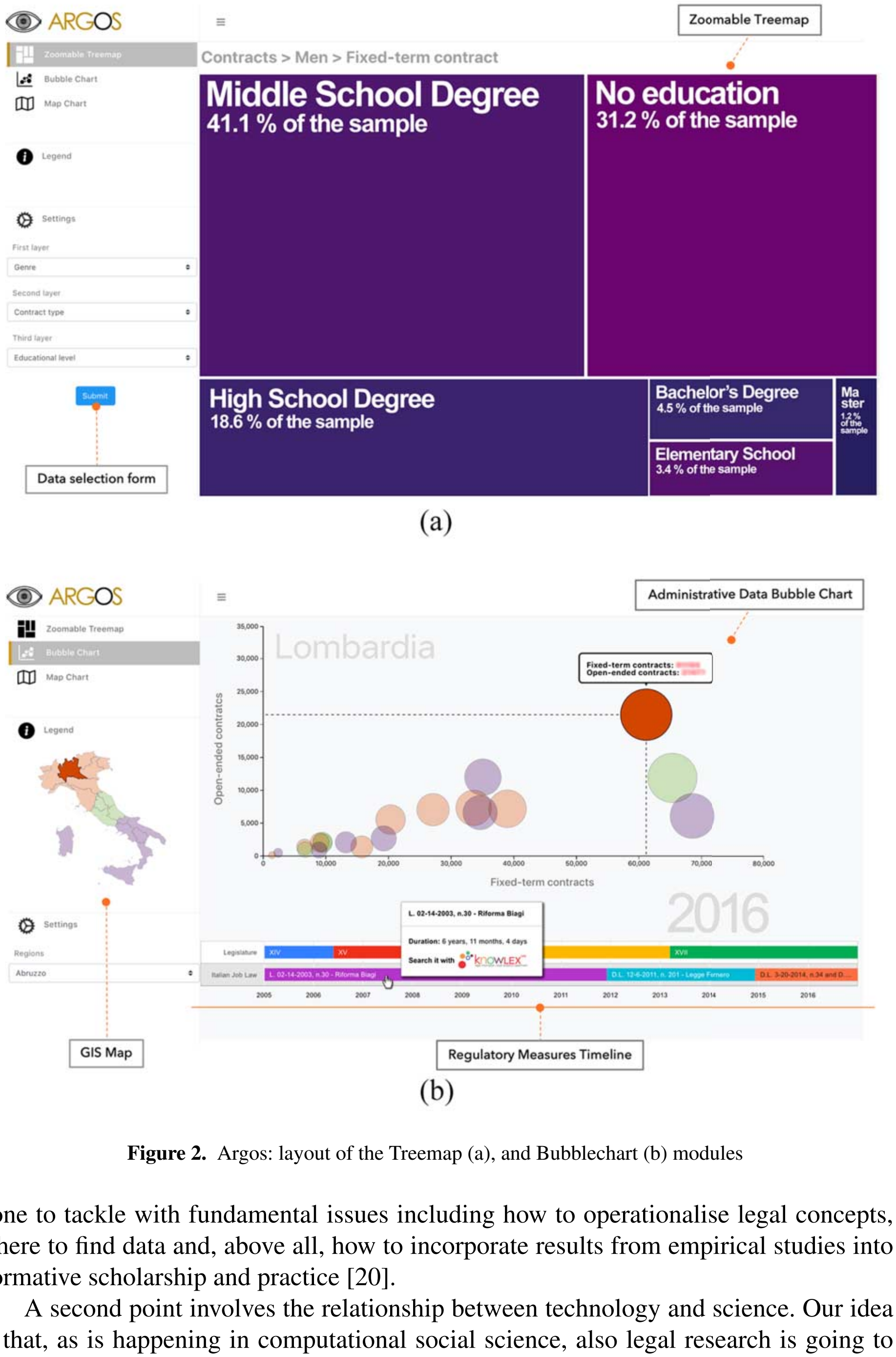
(2) Machine learning module
A second module, currently under development, aims at exploiting machine learning – in particular supervised learning algorithms [39] – and other techniques known as 'feature ranking' and 'feature selection' [40] to enhance our analysis. Last few years have been marked by a growing attention to the use of machine learning [38]; [39] in the analysis of economic data. This represents indeed an optimal solution when dealing with enormous amounts of data, when the data are gathered without a carefully controlled experimental design or when dealing with phenomena characterized by complicated nonlinear interactions. Against this backdrop, we have been (and still are) working to extract relevant information from administrative data about workers and contracts aiming to answer questions like: "what are, among the many available, the characteristics that define the workers having more chances to succeed? Are there any patterns in the behaviour of firms and workers possibly linked to regulatory measures adopted over time? After some difficulties encountered in the management of the large quantities of records at our disposal the analysis has already led to interesting insights that will be soon made available online, always using interactive visualization solutions.

6. Conclusions: Four Critical Remarks

In light of the above, we can make some points dealing, in more general terms, with the potential impact of machine science heuristics and tools on the legal field. Experience gained throughout the aforementioned projects suggests that platform technology is not only the enabling factor of more efficient ways of extracting knowledge from legal information but also, potentially, the driver of more deep changes in legal research perspectives. The claim is worthy of a more in depth reflection.

A first remark deals with the very objects of legal research. Last decade has seen a growing attention towards the empirical study of law also by means of quantitative and data-driven approaches[8]. Legal scholars anyway have so far reserved little consideration to the possibility of drawing inspiration from new tools to reconsider more fundamental aspects of their research. We have the feeling that analytical platforms could open up the gates to a sort of 'computation-enhanced legal empiricism' [2], a kind of technologically-enhanced version of empirical legal research [18] exploiting computation not only to identify trends and correlations in case law but also to investigate new topics e.g. using social simulations to illuminate the intricate network of individual and social mechanisms through which law emerges, is applied, produces effects. Besides scientific spill-overs, the fostering of a more empirical stance towards legal world is beneficial also for practical reasons. Factual investigation of legal phenomena increasingly appears to be an indispensable condition for more effective legal solutions able to cope with the complexity of the real world. A great deal of work will obviously have to be

[8] In the overview of the Journal of Empirical Legal Studies it read "With the explosion in information technology, data sources on the legal system are improving in quality and accessibility. Compared with just a few years ago, researchers today can easily access original data sets. [. . .] The time is ripe for empirical studies of the legal system".



(a)

First layer: Genre

Second layer: Contract type

Third layer: Educational level

Contracts > Men > Fixed-term contract

Regions: Abruzzo

L. 02-14-2003, n.30 - Riforma Biagi

Duration: 6 years, 11 months, 4 days

(b)

Figure 2. Argos: layout of the Treemap (a), and Bubblechart (b) modules

done to tackle with fundamental issues including how to operationalise legal concepts, where to find data and, above all, how to incorporate results from empirical studies into normative scholarship and practice [20].

A second point involves the relationship between technology and science. Our idea is that, as is happening in computational social science, also legal research is going to become more and more an 'instrument-enabled' practice [41]. Hence lawyers will soon be facing the design of new *ad hoc* 'machines' and heuristics if they want to advance their understanding of legal and social phenomena. The shift, it should also be stressed, is full of epistemological implications of which it is good to be aware. The way we design and use research tool is a theory-laden process and this is true also for information technologies and analytical platforms. Decisions about data to be processed, functionalities, analytical methods to be implemented in the tool incorporate basic scientific options as well as, just to give an example, the choice to use agent-based simulation models underpins a generative and micro foundational approach to the study social phenomena. Scientific perspectives, research findings and methods coevolve with research instruments. Legal scholars engaged in the design of computational tools will have therefore to carefully dwell on the impact of the new heuristics on their research questions, conceptual categories, methods of study and work.

Our third consideration touches on methodological issues. In social science, the idea of overcoming what has been called the *war of paradigms* [42], has gradually led to the emergence of a pluralist perspective [43]; [44] according to which the integration of different research methods [45] is crucial to enhance the understanding of social phenomena. This is turning to be real not only in the more traditional areas of social research, but also in emerging fields like computational social science [41] and computational legal research where the merger of heterogeneous research methods spanning from data mining to social simulation or network analysis is ever more frequent. Thanks to the technical sophistication of the tools, and to the high levels of interoperability between applications achieved by web technologies, analytical platforms represent today the ideal place for the integration of different research methodologies.

Last remark, somehow extending previous points, deals with the potential role of analytical platforms in promoting the adoption of more interdisciplinary approaches to research, an issue that appears to be growingly topical also in the legal field (see, for an overview, [46]). In recent years [47] interdisciplinarity has been given increasing attention being seen not only as a scientific option, but also as an obligatory step to manage complex and pressing real world issues that “cannot be adequately addressed by people from just one discipline” [48]. The statement fits very well our case: giving an answer to many questions of legal science and practice – assess the impact of legal norms; understand the deep nature of legal systems; predict the evolution of law enforcement strategies – is a complex task involving scientific knowledge that transcend the boundaries of traditional legal scholarship. Our ability to integrate in new ways different knowledge and disciplines becomes therefore crucial and analytical platforms can play a relevant role to this end. Similarly to the computer-based artefacts conjured up by Parisi in [49], they provide scientist belonging to different research areas with powerful toolkits to develop integrated and non-disciplinary analyses of complex phenomena.

In an highly-cited paper dwelling on the lack of scientificity of legal scholarship [50], Richard Posner advocated a new approach to the study of law using the methods of scientific inquiry to more deeply understand legal systems. Taking as reference “the prestige and authority of scientific and other exact modes of inquiry in general” – among which he explicitly mentions those coming from computer science – Posner calls for a more prominent role of science in legal world and for a more interdisciplinary research attitude seen as an essential condition for the “understanding and improvement of the legal system”. The change will probably take time and a gradual process of cross-fertilisation allowing to modify theoretical and methodological constructs entrenched over time. Online analytical platforms are certainly not the only means to trigger this transition, but will for sure play a role in finding new ways to generate legal knowledge in the Big data era. What it takes is the capacity to creatively merge different perspectives and also the daring to concretely experiment new scientific and methodological solutions, as tough as that might be. The issue is not to believe in utopias, but to build prototypes.

Acknowledgements

Ideas proposed in this work largely stem from a longstanding interdisciplinary collaboration with the colleagues and friends of the Department of Computer Science of the University of Salerno. The Author wishes to thank Delfina Malandrino, Rocco Zaccagnino, Alfonso Guarino, Pietro Russo, Antonio Altamura, Armando Faggiano, Luca Vicidomini. Author is also truly grateful to Margherita Vestoso for the insightful comments on the content and the title of this work as well as for the support given in proofreading the paper.

References

[1] Bratton, B. H. (2016). *The Stack: On Software and Sovereignty*. MIT Press.

[2] Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

[3] Evans, J. & Rzhetsky, A. (2010). Machine Science. *Science, 329*(5990), 399-400.

[4] Reed, D. A., Bajcsy, R., Fernandez, M. A. et al. (2005). *Computational Science: Ensuring America's Competitiveness*. President's Information Technology Advisory Committee Arlington VA.

[5] Hey, T., Tansley, S. & Tolle, K. M. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft research, Vol. 1.

[6] Boulton, G., Campbell, P., Collins et al. (2012). Science As an Open Enterprise. *The Royal Society.*

[7] Hey, T. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. C. & Hey, A. J. G. (Eds.). *Grid Computing: Making the Global Infrastructure a Reality*. Wiley and Sons, 809-824.

[8] Hine, C. (Ed.) (2006). *New Infrastructures for Knowledge Production: Understanding e-Science*. IGI Global.

[9] Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine, 16*(7), http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

[10] Venkatraman, V. (2013). When All Science Becomes Data Science. *Science*, https://www.sciencemag.org/careers/2013/05/when-all-science-becomes-data-science.

[11] Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big data & society, 1*(1).

[12] Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese, 169*(3), 615-626.

[13] Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.

[14] Crouzier, T. (2017), *Digital Tools for Researchers*, http://connectedresearchers.com/online-tools-forresearchers/.

[15] Katz, D. M., Bommarito II, M. J. & Blackman, J. (2017). A General Approach for Predicting the Behavior of the Supreme Court of the United States. *PloS one, 12*(4).

[16] Katz, D. M. & Bommarito, M. J. (2014). Measuring the Complexity of the Law: The United States Code. *Artificial Intelligence and Law, 22*(4), 337-374.

[17] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

[18] Cane, P. & Kritzer, H. M. (2012). *The Oxford Handbook of Empirical Legal Research.* Oxford University Press.

[19] Smits, J. M. (2012). *The Mind and Method of the Legal Academic*. Edward Elgar Publishing.

[20] Leeuw, F. L. & Schmeets, H. (2016). *Empirical Legal Research: A Guidance Book for Lawyers, Legislators and Regulators*. Edward Elgar Publishing.

[21] Faro, S. & Lettieri, N. (2013). *Law and Computational Social Science*. Edizioni Scientifiche Italiane.

[22] Ruhl, J., Katz, D. M. & Bommarito, M. J. (2017). Harnessing Legal Complexity. *Science, 355*(6332), 1377-1378.

[23] Russell, H. & Susskind, R. (2013). Tomorrow's Lawyers: An Introduction to Your Future. *Legal Information Management, 13*(4), 287-288.

[24] Berk, R. A., Sorenson, S. B. & Barnes, G. (2016). Forecasting Domestic Violence: A Machine Learning Approach To Help Inform Arraignment Decisions. *Journal of Empirical Legal Studies, 13*(1), 94-115.

[25] Lee, K. J., Azyndar, S. & Mattson, I. A. (2015). A New Era: Integrating Today's Next Gen Research Tools Ravel and Casetext in the Law School Classroom. *Rutgers Computer & Technology Law Journal, 41*, 31.

[26] Kuppevelt, D. & van Dijck, G. (2017). Answering Legal Research Questions About Dutch Case Law with Network Analysis and Visualization. *Legal Knowledge and Information Systems, 95*.

[27] Ourednik, A., Nellen, S. & Fleer, P. (2016). *A Visual Approach to the History of Swiss Federal Law*, http://www.dhd2016.de/abstracts/vortr%C3%A4ge-047.html

[28] Xu, J. J. & Chen, H. (2005). Crimenet Explorer: A Framework for Criminal Network Knowledge Discovery. *ACM Transactions on Information Systems, 23*(2), 201-226.

[29] Hutchins, C. E. & Benham-Hutchins, M. (2010). Hiding in Plain Sight: Criminal Network Analysis. *Computational and Mathematical Organization Theory, 16*(1), 89-111.

[30] Ferrara, E., Meo, P. D., Catanese, S. & Fiumara, G. (2014). Detecting Criminal Organizations in Mobile Phone Networks. *Expert Systems with Applications, 41*(13), 5733-5750.

[31] Didimo, W., Liotta, G. & Montecchiani, F. (2014). Network Visualization for Financial Crime Detection. *Journal of Visual Languages and Computing, 25*(4), 433-451.

[32] Taha, K. & Yoo, P. D. (2016). SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization. *IEEE Transactions on Information Forensics and Security, 11*(4), 811-822.

[33] Lettieri, N., Altamura, A. & Malandrino, D. (2017). The Legal Macroscope: Experimenting with Visual Legal Analytics. *Information Visualization, 16*(4), 332-345.

[34] Lettieri, N., Altamura, A., Faggiano, A. & Malandrino, D. (2016). A Computational Approach for the Experimental Study of EU Case Law: Analysis and Implementation. *Social Network Analysis and Mining, 6*(1), 56.

[35] Lettieri, N., Malandrino, D. & Vicidomini, L. (2017). By Investigation, I Mean Computation. *Trends in Organized Crime, 20*(1-2), 31-54.

[36] Lettieri, N., Altamura, A., Malandrino, D. & Punzo, V. (2017). Agents Shaping Networks Shaping Agents: Integrating Social Network Analysis and Agent-Based Modeling in Computational Crime Research. In Oliveira, E., Gama, J., Vale, Z., Cardoso, E. L. (Eds.). *Progress in Artificial Intelligence. Proceedings of the 18th EPIA Conference on Artificial Intelligence*. Springer, 15-27.

[37] Lathrop, D. & Ruma, L. (2010). *Open Government: Collaboration, Transparency, and Participation in Practice*. O'Reilly Media.

[38] Einav, L. & Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy, 14*(1), 1-24.

[39] Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

[40] Novaković, J. (2016). Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms. *Yugoslav Journal of Operations Research, 21*(1).

[41] Cioffi-Revilla, C. (2014). *Introduction to Computational Social Science*. Springer.

[42] Eckstein, H. (1998). Unfinished Business: Reflections on the Scope of Comparative Politics. *Comparative Political Studies, 31*, 505-534.

[43] Della Porta, D. & Keating, M. (2008). *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective*. Cambridge University Press.

[44] Sil, R. & Katzenstein, P.J. (2010). *Beyond Paradigms: Analytic Eclecticism in the Study of World Politics*. Palgrave Macmillan.

[45] Teddlie, C. & Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage.

[46] Siems, M. M. (2009). The Taxonomy of Interdisciplinary Legal Research: Finding the Way Out of the Desert. *Journal of Commonwealth Law and Legal Education, 7*(1), 5-17.

[47] Frodeman, R., Klein, J. T. & Pacheco, R. C. D. S. (2017). *The Oxford Handbook of Interdisciplinarity*. Oxford University Press.

[48] Nature (2015). *Mind Meld: Interdisciplinary Science Must Break Down Barriers Between Fields To Build Common Ground*, https://www.nature.com/news/mind-meld-1.18353.

[49] Parisi, D. (2014). *Future Robots: Towards a Robotic Science of Human Beings*. John Benjamins Publishing Company.

[50] Posner, R. A. (1986). The Decline of Law As an Autonomous Discipline: 1962-1987. *Harvard Law Review, 100*, 761.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190005

Entropy in Digital Information and the Enforcement of Law: Towards a Unification of Remedies?

Giorgio GIANNONE CODIGLIONE
Università degli Studi di Salerno (Italy)

Abstract. The Chapter aims to examine the legal remedies – both judicial and non-judicial – available in the area of electronic communication, adopting as the main comparison parameter the problem of the legal status of digital information. The infocentric structure of today's society on the one hand does not allow for the advance identification of a clear and generalized correspondence between a subjective legal situation and digital information; on the other hand, protection mechanisms tend to converge both from a classification and a technical profile. In other words, the consolidated subjective right vs. remedy model – understood as a system of subjective situations that are pre-established by the law from which owners derive their faculty or powers and which puts the obligation to do (or also not do) in the hands of individuals or the rest of the community, and alongside which a range of protection instruments can be found that can be invoked before the courts in the case of violations (*ubi jus*, *ibi remedium*) – is often diminished and becomes more typically an action-reaction model. In a multi-subject context marked by a post-industrial, cognitive economic model, it is possible that at the operational level the administration of one type of remedy implies a different consequence for all the other subjects involved in the information flow. While respecting the diversity of the experiences analysed, the regulatory trend seems to be that of the parcelling up of behavioral standards in a preventive and collaborative key.

Keywords. theory of information, economic analysis of law, defamation, data protection, intellectual property, consumer protection, tort law, remedies

1. Introduction

Tackling the subject of enforcement of the law on the Internet leads us firstly to highlight three fundamental differences between the online and offline perspective:

1. As we all know, a person connects to the Internet through one or more terminals with the general expectation of anonymity (protected or unprotected) [1]; [2]; [3]; [4]; [5]; [6]; [7]; [8]; [9].
2. The access to the Internet and its services generates a legal connection of interdependence with the network and service providers and produces a constant flow of data, characterised by the endless possibilities of using in terms of replicability, modifications and ubiquity.
3. Data that transits over the Internet can identify the agent, but also the action: the creation of a technical ecosystem that produces and feeds itself with digital in-

formation makes the person as a ‘complex information entity’ [10], that acts ever more often through digital inputs, establishing a two-way relationship of production and sustenance.

In other words, data represents the most innovative form of social subject ever created by man [11]; [12]; [13], that condenses (also simultaneously) the functions of identification and actions, just as it does those of prescriptive, sanctionatory or reparatory functions of a norm or a binding legal decision [14]; [15].

The ‘polisemic’ nature of the legal meanings and the effects attributed to online data, where this is not associated with a continuous and compliant flow of information, brings up the problem of entropy as developed in thermodynamics, recalled by Norbert Wiener [16], the father of cybernetics and later used in many areas of the modern theories on information, even though this needs to be contextualised within the particular system of the Internet.

Physicists define entropy as the measure of disorder in a system, hence the more ordered and structured the system is, the less entropy (and the more information) there is and vice versa. In a wider sense, the concept of entropy associates disorder and uncertainty inherent in the natural state[1].

The Internet, understood as a decentralised communication system that encourages technical and human interaction through the communication of data packages from one point to another[2] [17]; [18]; [19], represents in theory an exceptional tool to counter the phenomena of entropy, but at the same time unveils new risks of degeneration typical of technology as the maximum expression of man’s will for power [20]; [21]; [22]; [23]. According to Wiener, in fact, the main tool to control entropy is that of feedback, i.e. the human approach to act bearing in mind that which he has done before, thereby repairing any mistakes made[3].

On the Internet, this remedial function is in theory encouraged by the ease with which it is possible to communicate between different parties[4], but at the same time it suffers from the multiplier effect generated by the simultaneous and immediate circulation of non-compliant data, making it much more difficult to correct and modify the erroneous information within the system before they can have an effect[5].

Understanding these essential aspects of human actions on the Internet in more depth leads to a silent, but revolutionary, paradigm shift in the techniques of protection of the fundamental rights, requiring a holistic and multilevel approach, paying attention to verify how the attribution of a particular ‘legal statute’ to digital information can influence the different levels of the technological ecosystem [24].

2. Techniques of Protection of Online Reputation

Here I will try to outline two regulatory tendencies that are emerging – above all in Europe – and that concern three key areas of ICT Law: online defamation, data protection and copyright.

As is well-known, information (be it either digital or analogical) is considered as a ‘good’ [25]; [26]; [27]; [28]; [29]; [30]; [31], in the sense that it undergoes a process of ‘cataloguing’ in order to assign an owner and attribute a legal ‘value’ to it (which could also correspond in various measure also, ex ante or ex post, to an economic value) as well as the predisposition of a certain protection regime, consisting, for example, of the fixing of certain rules of conduct and the guarantee of the effectiveness of those remedies. Every definition of a certain ‘good’ has three functions: recognise the owner, attribute a value, act either before and/or after as a protection.

In the area of digital information on the Internet, due to the reasons outlined previously, such functions often seem to be limited, above all regarding the preventative definition in the legal system of the relative areas of protection [32].

Reputation is an information ‘good’ that, despite being recognised *a priori* by the legal system, is often evaluated in a negative sense, in that it is identified as such every time is harmed.

The creation and the posting of false, inexact, or misleading data regarding a natural or legal person by third parties on the Web, is one of the crucial issues of the first thirty years of the history of the Internet. The infinite replication and the ease with which harmful and defamatory content is circulated has led legislators, judges (and also ISPs) to apply two different protection methods, that in general can be defined following the distinction established by Calabresi and Melamed between *property rules* and *liability rules*, but with the particularity that I will try to explain now.

Property rules refer to the rules that protect in an absolute and exclusive way – for example through a prohibition or an order – the owner of a good from any form of interference from third parties, except in cases in which the parties decide voluntarily to cede that good to a third party in exchange for payment [33].

In our case, reputation is a subjective situation (but also a fundamental right of the individual) that must not be harmed in any way by other members, except in the context of the balancing of other rights and fundamental freedoms such as that of the freedom of expression.

The prohibition to offend or unjustifiably harm the image of another person, once violated, is evaluated and sanctioned with the removal of the harmful content, the order not to repeat this form of conduct and the payment of a sum of money to compensate the victim for the damage suffered. In the latter case, therefore, where the legislation retains it difficult (or uneconomical) to enact (and to enforce) a property rule (hence in the presence of high transaction costs), it can cumulatively, or alternatively, sanction the transgressor by imposing the *liability rule*, defined by Calabresi and Melamed as a remedy to compensate the victim in cases where it is impossible or uneconomical for the person doing the harm and the person harmed to find an agreement for the transfer of the good.

On the so-called Web 2.0, the protection framework can be summarised as follows: given that the user must be able to operate freely on the Internet, in the case of content created and circulated on platforms managed by third parties, it is the provider that is required to act – on the input of the judicial authority or by the individual who has been harmed – to supply the identification data of the author of the unlawful content and/or remove the harmful information, or also to respond for the eventual failure to remove this content in a timely manner [34].

In other words, the system is based on the general activation of the property rule connected to the degree of collaboration of the provider (supply of the data regarding the author of the conduct, timely removal of the content). In essence, due to the difficulty in tracing the author of the damage, which is linked to the protection of the legitimate expectation of privacy of the user, the protection regime was characterised by the minimum goal of the removal of the content, to which is linked the eventual activation of the liability rule – such as sanctions for the failure (or incorrect) fulfilment of a supervision duty given to the provider and consistent with the *ex post* protection of the rights of the users.

The framework described, proposed in the Electronic Commerce Directive 2000/31/EC and outlined in a more liberal way in the *U.S. Common Decency Act of 1996*, in the last decade has had to face the problem of the implementation of systems of social interaction that are entirely based (not only because of how they function, but also their economic sustainability) on the incentive for users to produce and share data.

The user is in fact encouraged to become an integral part of the system and this leads to the difficulty in controlling the flow of information and erasing harmful content in a definitive manner. This is a problem that is added to that already described of the identification of the true author of the conduct.

A recent judicial trend has strengthened the rule of the direct responsibility of the user, that places on the person who created and circulated the harmful content not only the duty to remove it, but also to act in order to prevent any form of replication or diffusion of the information.

The *Störerhaftung* rule, connected to an interpretation of § 1004 BGB (regarding the actions taken to remove the content that violates copyright or to inhibit its continuation) is aimed at all those who have either directly (*unmittelbarer Störer*) or even as an intermediary or indirectly (*mittelbarer Störer*) caused an unauthorised invasion into the judicial sphere of another[6], confirming a rule of responsibility that is particularly sensitive to the causal relationship of the harmful conduct, also in the presence of the extreme fragmentation induced by the presence of the user in the technical-informational ecosystem.

The *Störerhaftung* is set out as a property rule that is extended by case law also to the categories of the Individual - *Immaterialgüter rechte* and is applied if the ‘interferer’: a) contributed to the illegal activity in a causal and inappropriate manner; b) had the legal and practical possibility to prevent the violation; c) violated the reasonable duty of supervision or control over such illegal conduct for prevention purposes. In general, the violation of these rules does not lead to a sentence for compensation damages, but at most the payment of legal costs for any notification of cease and desist orders[7].

The responsibility linked to the possible circulation of content by third parties is partially given also to the person who initiated the communication: they must in fact take all adequate, proportionate and not excessively costly measures – with the cooperation of other users (who are in turn obliged to do as *Täter*, i.e. direct disturbers) – in order for the harmful content to be removed entirely from the Internet.

As for the rules regarding the duty of the ISP, a new direct policy tendency seems to be emerging, confirmed – as we will see – by some adaptations of case law[8].

Following the principle of neutrality confirmed in Article 15(1) of the Directive 2000/31/EC, providers do not have an overall duty to supervise the information that they transmit or store, or any obligation to actively look for facts and circumstances that indicate the presence of unlawful activity. This assumption, that has been reiterated many times by the EU Court of Justice and that became one of the key turning points in the evolutionary judicial interpretation of the e-commerce directive [35], finds a new interpretation in the recent position announced by the EU Commission[9].

According to the Commission, the heterogeneity of the regulatory approaches implemented over the years by the Member States, differentiated on the basis of the procedure and often dedicated to specific types of unlawful information, gives rise to a fragmented and ineffective framework.

From this standpoint, EU strategy begins with the promotion of voluntary measures (such as the *Code of Conduct to combat illegal forms of online hate speech*[10]) and others focused on the fight against specific crimes (such as child pornography or terrorism[11]), until reaching the provision of a dual channel of measures that strengthen the level of diligence required from hosting providers.

In terms of prevention, a discussion is underway on the adoption of effective proactive voluntary measures, aimed at identifying and removing illicit contents in order to reduce the risk of serious damage, without however the need for the provider to move outside the regime of applicability of the safe harbour regime referred to in Article 14, Directive 2000/31/EC[12].

In subsequent interventions, the Commission calls for the implementation of easily accessible and understandable mechanisms for the gathering of motivated and precise reports, which also guarantee the timely and effective removal or disabling of access to illegal content following reports from users[13].

3. The EU's Data Protection Strategy and Its Influence on the Evolution of ICT Law

The picture described so far foresees, therefore, the existence of a legal relationship – corresponding to different forms of supervision duty – between the user and the information introduced and circulated by them on a web platform. This tendency, interpreted backwards, reallocates the property rule in a greater measure to the author (or the co-authors) of the damage, framing the intervention rules and the responsibilities of the provider as a form of secondary regime, that is in turn split into prevention and reparation duties.

Also in the area of the protection of personal data, the regulatory tendencies seem to follow this new viewpoint, strictly linked to the parcelization of conduct with the creation of a qualified relationship between information and those involved in the processing activity.

The fundamental right to data protection is understood as a subjective legal situation that ascribes abstractly the 'belonging' to natural person of all the data that is directly or indirectly attributable to them. In the absence of consent and/or of a justifying cause, the processing of any information concerning an identified or identifiable natural person is illicit and leads to consequences for the parties who are involved in the processing to varying degrees (the controller, the joint controller, the processor, those authorised to process data).

Given the fundamental role that the circulation of data plays in the Internet environment and more generally in today society, Convention no. 101/1981 of the Council of Europe and Directive 95/46/EC (first) and General Data Protection Regulation 679/2016 (GDPR), followed in 2018 by the CoE's 'Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data' (after), established the evolution of the right to data protection on one hand as a right to the control and to the information of the data subject (*Recht auf informationelle selbstbestimmung*), and on the other hand, as an obligation for compliant processing for controllers and processors, establishing precautionary conduct and introducing a harsh administrative sanctions regime that in turn compensates for the mitigation of the criminal and civil sanctions regime.

In other words, the parties that undertake a processing activity[14] must adhere to strict rules in order not to suffer – amongst the other sanctions – the possible imposition of a liability rule. The accountability principle introduced by GDPR consists in the duty to enact (and provide proof) organisational activities (e.g. impact assessment, pseudonymisation, privacy by default and by design) and preventive or specific redress remedies such as the modification, rectification and erasure of data, the limitation of processing or the opposition and the provision of personal data in an interoperable format (the so-called right to portability) [36]; [37]; [38].

The online 'right to be forgotten' is a clear example of the parcelization of the duties of care and also of the complementary position attributed to pecuniary compensation rules: following the jurisprudential adaptation undertaken by the CJEU with the *Google Spain* case[15], Article 17 of the GDPR affirms the autonomy of the processing carried out by web-pages with respect to the activity of search engines, establishing the right of the data subject to request directly to the data controllers respectively to evaluate respectively the possible deindexing or erasure of the information deemed inaccurate or processed in an unlawful manner. An analysis of the case law on the issue (in France, Italy and Spain) shows that failure to properly comply with the screening of the requests does not necessarily lead to the obligation to pay damages, a measure that in most cases is limited to the refunding of legal and procedural costs[16].

As a remedy, paragraph 2 of the aforementioned Article 17 reformulates the rule proposed by the BGH on the issue of the protection of reputation: the data controller who made public the personal data considered illicit, must – "*taking into account the technology available and the costs of implementation*" – inform any other data controllers who are processing the same data of the request of the party concerned to delete any link, copy or reproduction of this[17]. If, for example, the editor of an online newspaper is obliged to delete some personal data contained in an article published in the past, they have the duty to inform also the other parties that supply further diffusion or links to that news story (as well as the search engines).

This trend should be read in the context of the interpretation of the concept of 'controller', 'joint controller' and 'processor' in a flexible and dynamic way not necessarily to the business figure of the ISP: these concepts can in fact be progressively extended for example to the user who manages a social profile not for mere domestic purposes, as established by recital no. 18 of the GDPR and confirmed by the CJEU in the recent case of *Wirtschaftsakademie Schleswig-Holstein* regarding the administration of a fan page[18].

4. Enforcing Digital Copyright and the Role of Internet Service Provider: From Reparation to Prevention?

Coming to digital copyright – understood as a right with a monopolistic-ownership matrix, having as its subject some particular forms of human creativity (such as literary works, musical works, but also databases) – it embodies the contrast par excellence between having the exclusive right over or having access to knowledge (and therefore the right to freely benefit from the information flow promoted by the Internet) [39]; [40]; [41].

Having lost the tangibility of the media used by intermediaries (such as publishers, producers, distributors), to 'capture' the intellectual work and facilitate the authorized circulation after the payment of a price by the users, digital copyright is constantly searching for a new structure that is capable of making the requirement to effective protect copyright holders without limiting the incentive goal that is generally recognized in the free circulation of creative works.

In this sense, on a level of taxonomy, case law and doctrine have for a decade ensured the transit of copyright from the dominical dimension to that of so-called compensation fees. However, this does not solve the problem of the effectiveness of the remedies that protect the patrimonial and non-patrimonial assets of the authors encapsulated in the digital information, through techniques aimed at controlling their circulation (think for example of the DRM, an expression of a property rule that is a fundamental part of the conformation technique), or through the prior identification of the subject to whom to allocate at least the 'price' of unauthorized use, where it is not 'protected' by the safe harbour of fair use.

In this framework, a new and interesting interpretation of the material has been given by a set of judgments of the CJEU.

In *GS Media vs. Sanoma*[19], the publication and updating of hyperlinks, repeated also following the removal of protected content by the source websites, was deemed to be an unauthorized act of communication to the public, as it was carried out for profit (e.g. related to the publication of advertising banners) and for this reason "*with full knowledge of the fact that the work is protected and that the copyright holder may not have authorized the publication on the Internet*"[20]. The mere fact that the sole activity of providing a connection to protected works published without authorization is undertaken by a party who performs a business activity, includes a presumption of knowledge of the unlawfulness of such contents.

Together with the "*unavoidable role of the user*" and the "*intentional nature of his/her intervention*", the reference to the "*profit-making purpose*", introduced for the first time in the Premier League decision[21] – and indeed absent in Article 3(1), Directive 2001/29/EC – is used by the Court as a further parameter for an 'individualized assessment' of the EU concept of 'communication to the public', ensuring a fair balancing of the public interest in access to information, recognized to the activity of creating links to third-party websites carried out by private individuals who are not working to obtain any profit. In fact, in the case of non-profit entities, the Court introduces a general regime of unawareness of the possible illegality of the content to which the link refers. This rule no longer applies once the user has received notice about this, in a manner similar to the provisions on the exemption from liability of the hosting provider in the e-commerce directive.

The presumption of the unlawfulness of linking related to the requirement of the awareness of the person who posts it, as well as the lucrative nature of the communication, has found confirmation in the *Stichting Brein I* case: the marketing of portable multimedia players equipped with a software capable of opening web links to pages in which protected works were made available without the authorization of the copyright holders, was deemed an act of unauthorized communication to the public, which did not meet the requirements of the exception on temporary reproduction referred to in Article 5(1), Directive 2001/29/EC, read in conjunction with Article 5(5)[22].

This trend has found confirmation in *Stichting Brein II*[23] (the so-called *Pirate Bay* case), in which the CJEU judged as an unauthorized 'communication to the public' the conduct of a web platform, which by indexing metadata related to protected works (so-called torrents) and the provision of a search engine allowed users to locate such works and share them over a network (*peer-to-peer*).

When interpreted together, the aforementioned decisions lead to a confirmation of a presumption of the responsibility of the provider who, for the pursuit of a profit, allows searching, location and sharing of protected works. In this case, the notion of profit is in any case implicitly linked to the fact that the provider offers a service that is addressed – entirely or in part – to the aggregation and provision of content uploaded by users (and in any case traceable to another physical address, represented by a website or even by the PC connected to the sharing network) and aimed at aware communicating to the public *en masse*, protected works.

The formulation of a general presumption of the knowledge of the unlawfulness of the contents in relation to the provider of links who acts for profit mainly, paves the way for the affirmation of a preventive duty of control placed also on generalist search engine [42].

To this, therefore, a parameter of imputation of semi-objective character should be applied, in some way comparable to that in force in the area of protection of personal data, given that the provider should be exempt from this only once this presumption has been confuted, by proving they "adopted all the appropriate measures to avoid the damage".

This direction is also confirmed by the recent approval by the European Parliament – albeit with some modifications – of the *Proposal for a Directive on copyright in the digital single market*[24], which in Article 13 foresees, for the 'new' category of online content sharing providers[25], the obligation to conclude licensing agreements with the owners of the rights of the works uploaded[26].

The proposal applies the obligation to negotiate licenses and to adopt content protection measures to a broad category of subjects, identifying in the "information society service providers one of the main purposes of which is to store and give access to the public or to stream significant amounts of copyright protected content uploaded/made available by its users, and that optimise content, and promote for profit making purposes, including amongst others displaying, tagging, curating, sequencing, the uploaded works or other subject-matter, irrespective of the means used, and therefore act in an active way".

In this way the provider – defined in the law in the category of the so-called 'active hosts' established through the interpretation of recital no. 42 of the Directive 2000/31/EC – is encouraged to act in a precautionary way to check and if necessary to remove the contents uploaded by users in order to avoid incurring liability. In other words, the proposal introduces a further supervision duty for the majority of Web 2.0 providers, who could thus be called on directly to respond to the failure (or incorrect) fulfilment of this obligation.

This supervision duty goes alongside that foreseen by the e-commerce directive, which, on the contrary, defines a regime of the general lack of responsibility of the provider, as long as they respect the precise rules of conduct (and timely intervention) in cases where they are notified of a violation [43].

The above mentioned proposal, now being examined by the Council, seems nevertheless to strengthen the idea of establishing an obligation of advance collaboration and negotiation of the rights of reproduction and/or communication to the public of protected works with the copyright holders or the assignees, linked to which is the implementation of systems to verify information uploaded and made accessible on platforms. This, in a certain sense leaves a complementary role to remedies in a specific and subsequent form such as injunctive measures or notice and take down procedures[27].

This legal policy option underpins a different classification of service providers from a legal-economical point of view: the rules of provider responsibility (or lack of) introduced between the end of the last century and the beginning of the new millennium are based on the need to incentivize the growth of information society services [44]; [45]; [46]; [47].

In this context, the construction of the new EU copyright seems to be based on the awareness of the existence of entrepreneurial positions (often a form of monopoly) that reach across different sectors of the culture industry previously managed exclusively by the traditional media, with the consequent imposition of 'costs' – in the case of copyright in the form of preventive obligations for the control and/or payment of fair remuneration – aimed at redistributing in part the wealth produced by the activity of intermediation and aggregation of contents, as well as allocating more efficiently the 'anonymous harm' resulting from the inherent difficulty on the Internet of attributing the responsibility of the illegal conduct directly to the true author.

It should also be noted that the interpretative trends described (especially in *GS Media* case, but also with reference to Article 13 of the Proposal) here consolidate the emergence of the hybrid figure of the so-called prosumer, obliterating the theories that see the user as a mere (non-paying) exploiter of data flows and leaving space to a vision that takes into due consideration the economic value of the inputs produced and supplied to providers from each access to or by surfing the Internet[28].

5. The Intersection Between the Personal and the Economic Dimension of Data

The profile analysed before is worthy of further consideration, before coming to some brief conclusions.

The Web 2.0 user does not play a solely passive role as a user of the services and content offered by providers, but is induced to implement data assets that circulates on the Internet primarily through the non-pecuniary exchange represented by their personal data, captured by the providers both on the open platforms (for example with cookies) and on closed ones, subject to a registration process which has a para-contractual nature.

Confirmation of this point emerges from the contribution of the legal formants of some Member States[29] and also from some legislative proposals that tend to evaluate some types of personal (and non personal) data produced by users as a form of payment[30]. In this way, we would recognize an exquisitely consumeristic scope of protection for a basic level of legal relations hitherto inadequately highlighted since masked behind the paradigm of free services[31].

This tendency – aimed at protecting the user's freedom of access to information also as an incentive to a 'participatory exploitation' of services [48] – seems to be implicitly confirmed by the results of the *GS Media* case, in which the user's position with respect to a violation of the copyright that occurred through the posting of a link to a third content is firstly protected by a presumption of not-awarness.

Instead, concerning the phase for calculating the possible sum of money due for compensation purposes – which is an essential step for assessing the concrete scope of a liability rule, it seems appropriate to outline three different trends, which are interrelated:

a) the importance and the topical nature of the specific redress remedy as a complementary measure to compensation, “where this is wholly or partially possible” (as envisaged for example in Italy by Article 2058 of the Civil Code)[32];

b) given the inherent nature of the process of data circulation, Tort liability is understood primarily as a tool for parcelling and controlling the duties of care on the Internet among all the subjects involved in various ways in the flow of information[33];

c) taking into consideration what is stated in *a*) and *b*) above, the phase of the computation of the sum due as compensation is closely linked to the importance of compliance with and prompt activation of the remedial measures, in accordance with the principle of non-excessive burdens for the debtor. The remaining sum must then be calculated in relation to variable parameters, for example depending on the type of information that is the subject of the harm and the nature of the damage suffered: whilst, on the one hand, the criterion of the ‘price of consent’ appears progressively to extend from IP rights to other spheres which by tradition are defined as highly personal [49]; [50], such as the right to personal identity and digital privacy, on the other hand the liquidation of non-pecuniary loss seems to preserve its deterrent-sanctioning dimension, balanced by a more rigorous system of evidence on the injured party.

6. Concluding Remarks

Coming to the first (and partial) conclusions, the relationship between digital information and the user on the Internet is distinguished by the particular degree of complexity, moving it away from the traditional models.

From the first point of view, all the parties (be they legal, physical persons or else robots – even though currently not assigned with their own personality by the legal system[34], have the ability/duty to come into contact with the data according to a principle of ‘digital solidarity’[35], ensuring that the flow of information is constant and that it is promptly corrected either as a precautionary or as a subsequent measure depending on supervision standards that vary according to the degree of proximity.

The owner of the right does not have an exclusive hold over the ‘good’ (whether it is the subject of a copyright or a right of the personality which is consolidated in their digital identity [51]; [52]; [53]), but releases this good into the information flow reserving for him/herself the right to observe the compliant use and the recognition of a flat-rate compensation fee that ‘perceives’ and ‘spends’ in the context of the permanence of the technical-informational ecosystem in either a contractual manner (e.g. through signing up for social media services and access to content available on this) or in an Aquilian manner (constituting harmful conduct).

The distinction between property rules and liability rules seems finally to fade, merging into a single remedial system. The elements collected so far propose in fact the image of a ‘weakened’ property rule in terms of scope and effectiveness (the remedy operate on a material level of the conduct and on the availability of unlawful data, but do not favour a negotiation based on the real value that the holder attributes to the right that is violated) and ‘disguised’ by a liability rule that by nature and entity is basically pre-established.

If we want to go further, the progressive ‘codification’ of online interactions (think e.g. of the advent of the blockchain technology) could open up the scenario of a definitive reduction in the para-contractual logic of relationships, especially in the cases in which they constitute pecuniary loss. The violation of the duties of supervision divided *pro quota* among the parties involved in the information flow may in fact lead, in a future of digital relations, to the automatic transfer of a sum of money (therefore connected to strict or aggravated liability), understood as an indemnity that integrates the effectiveness of remedies able to interrupt and eliminate the harm.

Given the progressive osmosis between private enforcement and judicial system, the notion of 'harm' appears ultimately to be represented as the phenomenon of the initial or repeated circulation of inaccurate or unauthorized information, in some way merging the specific redress functions (aimed at the fulfillment of subjective situations or to prevent the violation, almost entirely delegated to the intervention of providers and users) with those of a compensatory form (aimed at removing and repairing 'harm'). This scenario could for example obliterate one important goal attributed to Tort Law, that is the deterrence function [54]; [55] and more in general imposes a radical reflection on the consequences of the 'datafication', in the perspective of the progressive automation of all type of relationship (e.g.: human-human, human-machine, machine-machine, human-robot etc.).

Endnotes

[1] According to Wiener [16], p. 28, 40: "As we have said, nature's statistical tendency to disorder, the tendency for entropy to increase in isolated systems, is expressed in the second law of thermodynamics. We as human beings, are not isolated systems. (...) As entropy increases, the universe, and all closed systems in the universe, tend naturally to deteriorate and lose their distinctiveness, to move from the least to the most probable state, from a state of organization and differentiation in which distinctions and forms exist, to a state of chaos and sameness. In Gibbs' universe order is least probable, chaos most probable. But while the universe as a whole, tends to run down, there are local enclaves of whose direction seems opposed to that of the universe at large and in which there is a limited and temporary tendency for organization to increase. Life finds its home in these enclaves. It is with this point of view at its core that the new science of Cybernetics began its development".

[2] According to [19]: "In a system that includes communications, one usually draws a modular boundary around the communication subsystem and defines a firm interface between it and the rest of the system. When doing so, it becomes apparent that there is a list of functions each of which might be implemented in any of several ways: by the communication subsystem, by its client, as a joint venture, or perhaps redundantly, each doing its own version. In reasoning about this choice, the requirements of the application provide the basis for a class of arguments, which go as follows: The function in question can completely and correctly be implemented only with the knowledge and help of the application standing at the end points of the communication system".

[3] See [16], p. 26: "It is my thesis that the physical functioning of the living individual and the operation of some of the newer communication machines are precisely parallel in their analogous attempts to control entropy through feedback. Both of them have sensory receptors as one stage in their cycle of operation: that is, in both of them there exists a special apparatus for collecting information from the outer world at low energy levels, and for making it available in the operation of the individual or of the machine. In both cases these external messages are not taken neat, but through the internal transforming powers of the apparatus, whether it be alive or dead. The information is then turned into a new form available for the further stages of performance. In both the animal and the machine this performance is made to be effective on the outer world. In both of them, their performed action on the outer world, and not merely their intended action, is reported back to the central regulatory apparatus. This complex of behavior is ignored by the average man, and in particular does not play the role that it should in our habitual analysis of society; for just as individual physical responses may be seen from this point of view, so may the organic responses of society itself. I do not mean that the sociologist is unaware of the existence and complex nature of communications in society, but until recently he has tended to overlook the extent to which they are the cement which binds its fabric together".

[4] *Ibidem*, p. 31: "Just as entropy is a measure of disorganization, the information carried by a set of messages is a measure of organization. In fact, it is possible to interpret the information carried by a message as essentially the negative of its entropy, and the negative logarithm of its probability. That is, the more probable the message, the less information it gives. Cliches, for example, are less illuminating than great poems".

[5] According to [12], p. 130: "*Nous venons de dire que la Technique produit au profit de l'homme des compensations aux inconvénients, qu'elle se produit pour elle-même des facilitations, et peut changer de caractère (décentralisation) cependant, il apparaît de plus en plus que ce système ne possède actuellement pas une des caractéristiques considérées généralement comme essentielle pour un système: le feedback, la rétroaction c'est-à-dire, rappelons-le d'un mot, ce mécanisme qui intervient lors- qu'un ensemble, un système en mouvement commet une erreur dans son fonctionnement, pour rectifier cette erreur mais en agissant à la source, à l'origine du mouvement. Il n'y a pas 'réparation' de l'erreur commise, il y a reprise du mouvement à son origine en modifiant une donnée du système. Le feedback n'existe pas seulement dans les systèmes mécaniques, artificiels, mais aussi dans les systèmes biologiques ou écologiques. Il implique un contrôle des résultats suivi d'une rectification du processus lorsque les résultats contrôlés sont nocifs ou insatisfaisants. Ainsi le système technique ne tend pas à se modifier lui même lorsqu'il développe des encombrements, des nuisances, etc., il est livré à une croissance pure, dès lors ce système provoque un accroissement des irrationalités, et d'autre part, il est d'une lourdeur et d'une viscosité considérable: lorsque l'on constate des désordres et des irrationalités, cela n'entraîne rien que des processus compensatoires. Le système continue à évoluer dans sa propre ligne*".

[6] BGH, 28 July 2015 - VI ZR 340/14. *Il Diritto dell'informazione e dell'informatica*, 2, 2016, 292, translation and comments by G. Eramo Puoti. On this matter, see also [56]; [57].

[7] See e.g. BGH, 12 May 2010 - I ZR 121/08, in *MIR*, 6, 2010; for further remarks [58]; [59]; [60].

[8] See e.g. the contribution of the European Court of Human Rights: ECtHR, Application no. 64569/09, Judgement 16 June 2015, *Delfi AS v. Estonia*; ECtHR, Application no. 22947/13, Judgement 2 May 2016, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*.

[9] EU Commission, *Tackling Illegal Content Online. Towards an Enhanced Responsibility of Online Platforms*, 28-09-2017, COM(2017) 555 final; Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, in OJ L 63/50. 6.3.2018.

[10] In 2016, four of the leading Web 2.0 service providers (Facebook, Twitter, YouTube and Microsoft, to which Instagram was added) in fact signed the Code of Conduct on countering illegal online hate speech, promoted by the Commission and non-governmental bodies, implementing measures for receiving and screening reports on contents non-compliant with the Framework Decision 2008/913/JHA of 28 November 2008, on combating certain forms and expressions of racism and xenophobia through the use of criminal law. In almost two years of implementation of the Code, the providers have intervened in the majority of cases removing the contents deemed unlawful under the Framework Decision and/or the laws of the Member States within 24 hours (about 70% of the reported contents).

[11] This reference is above all to Directive 2011/93/EU on the removal of child pornography websites or to Directive 2017/541 concerning the online contents that constitute incitement to carrying out crimes of terrorism.

[12] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, cit., recitals no. 25 and 26: "In addition to notice-and-action mechanisms, proportionate and specific proactive measures taken voluntarily by hosting service providers, including by using automated means in certain cases, can also be an important element in tackling illegal content online, without prejudice to Article 15(1) of Directive 2000/31/EC. In connection to such proactive measures, account should be taken of the situation of hosting service providers which, because of their size or the scale on which they operate, have only limited resources and expertise and of the need for effective and appropriate safeguards accompanying such measures. It can, in particular, be appropriate to take such proactive measures where the illegal character of the content has already been established or where the type of content is such that contextualisation is not essential. It can also depend on the nature, scale and purpose of the envisaged measures, the type of content at issue, on whether the content has been notified by law enforcement authorities or Europol and on whether action had already been taken in respect of the content because it is considered to be illegal content. With regard to child sexual abuse material in particular, hosting service providers should take proactive measures to detect and prevent the dissemination of such material, in line with the commitments undertaken in the context of the Global Alliance against Child Sexual Abuse Online". These preventive obligations are therefore placed within the scope of applicability of Recital no. 48, Directive 2000/31/EC, being able to affect the safe harbor of the provider in the event of failure to intervene. On the one hand, in fact, the Commission excludes that by adopting these measures, the provider can be automatically be placed alongside the active hosting providers. (EU Commission, *Tackling Illegal Content Online*, cit., 12: "the mere fact that an online platform takes certain

measures relating to the provision of its services in a general manner does not necessarily mean that it plays an active role in respect of the individual content items it stores and that the online platform cannot benefit from the liability exemption for that reason"). On the other hand, it states that the platform, once aware of the illegality of the content during the course of preventive control activities, could lose its immunity if it does not act promptly to remove or disable access to illicit information (*Ibidem*, 13: "It follows that proactive measures taken by an online platform to detect and remove illegal content may result in that platform obtaining knowledge or awareness of illegal activities or illegal information, which could thus lead to the loss of the liability exemption in accordance with point (a) of Article 14(1) of the E-Commerce Directive. However, in such cases the online platform continues to have the possibility to act expeditiously to remove or to disable access to the information in question upon obtaining such knowledge or awareness. Where it does so, the online platform continues to benefit from the liability exemption pursuant to point (b) of Article 14(1). Therefore, concerns related to losing the benefit of the liability exemption should not deter or preclude the application of the effective proactive voluntary measures that this Communication seeks to encourage").

[13] EU Commission, *Tackling Illegal Content Online*, cit., 8: "Given that fast removal of illegal material is often essential in order to limit wider dissemination and harm, online platforms should also be able to take swift decisions as regards possible actions with respect to illegal content online without being required to do so on the basis of a court order or administrative decision, especially where a law enforcement authority identifies and informs them of allegedly illegal content. At the same time, online platforms should put in place adequate safeguards when giving effect to their responsibilities in this regard, in order to guarantee users' right of effective remedy".

[14] According to Article 4, no. 2) of GDPR, 'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

[15] CJEU, grand chamber, 13 May 2014, C-131/12, *Google Spain, Google Inc. v. AEPD, Costeja González.*

[16] See TGI Paris, 24 November and 19 December 2014, Marie-France M. v. Google France e Google Inc. *Il Diritto dell'informazione e dell'informatica*, 2015, 532; Trib. Roma, 3 December 2015, X v. Google Inc., *ivi*, 2016, 266; Sala de lo Contencioso-administrativo de la Audiencia Nacional, 29 December 2014, no. ric. 725/2010, 18 persone v. Google Spain e Google Inc., www.poderjudicial.es.

[17] Similarly, Article 19 GDPR on 'Notification obligation regarding rectification or erasure of personal data or restriction of processing' states that: "The controller shall communicate any rectification or erasure of personal data or restriction of processing carried out in accordance with Article 16, Article 17(1) and Article 18 to each recipient to whom the personal data have been disclosed, unless this proves impossible or involves disproportionate effort. The controller shall inform the data subject about those recipients if the data subject requests it".

[18] CJEU, grand chamber, 5 June 2018, C-210/16, *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH*, paras. 33-43: "(...) Moreover, other entities such as Facebook partners or even third parties 'may use cookies on the Facebook services to provide services [directly to that social network] and the businesses that advertise on Facebook'. That processing of personal data is intended in particular to enable Facebook to improve its system of advertising transmitted via its network, and to enable the fan page administrator to obtain statistics produced by Facebook from the visits to the page, for the purposes of managing the promotion of its activity, making it aware, for example, of the profile of the visitors who like its fan page or use its applications, so that it can offer them more relevant content and develop functionalities likely to be of more interest to them. While the mere fact of making use of a social network such as Facebook does not make a Facebook user a controller jointly responsible for the processing of personal data by that network, it must be stated, on the other hand, that the administrator of a fan page hosted on Facebook, by creating such a page, gives Facebook the opportunity to place cookies on the computer or other device of a person visiting its fan page, whether or not that person has a Facebook account. (...) In those circumstances, the administrator of a fan page hosted on Facebook, such as Wirtschaftsakademie, must be regarded as taking part, by its definition of parameters depending in particular on its target audience and the objectives of managing and promoting its activities, in the determination of the purposes and means of processing the personal data of the visitors to its fan page. The administrator must therefore be categorised, in the present case, as a controller responsible for that processing within the European Union, jointly with Facebook Ireland, within the meaning of Article 2(d) of Directive 95/46. (...) In those circumstances, the recognition of joint responsibility of the operator of the social network and the administrator of a fan page hosted on that network in relation to the process-

ing of the personal data of visitors to that page contributes to ensuring more complete protection of the rights of persons visiting a fan page, in accordance with the requirements of Directive 95/46. However, it should be pointed out (...) that the existence of joint responsibility does not necessarily imply equal responsibility of the various operators involved in the processing of personal data. On the contrary, those operators may be involved at different stages of that processing of personal data and to different degrees, so that the level of responsibility of each of them must be assessed with regard to all the relevant circumstances of the particular case".

[19] CJEU, 8 September 2016, C-160/15, *GS Media BV v. Sanoma Media Netherlands BV, Playboy Enterprises International Inc., Britt Geertruida Dekker*, in particular paras. 45 and 54.

[20] CJEU, C-160/15, cit., par. 50.

[21] CJEU, grand chamber, 4 October 2011, C-429/08, C-403/08 and C-429/08, *Football Association Premier League Ltd. and others*.

[22] CJEU, 26 April 2017, C-527/15, *Stichting Brein v. Jack Frederik Wullems*.

[23] CJEU, 14 June 2017, C-610/15, *Stichting Brein v. Ziggo BV, XS4ALL Internet BV, (Stichting Brein II)*.

[24] COM(2016) 593 final.

[25] See Recital no. 37a: "Certain information society services, as part of their normal use, are designed to give access to the public to copyright protected content or other subject-matter uploaded by their users. The definition of an online content sharing service provider under this Directive shall cover information society service providers one of the main purposes of which is to store and give access to the public or to stream significant amounts of copyright protected content uploaded / made available by its users, and that optimise content, and promote for profit making purposes, including amongst others displaying, tagging, curating, sequencing, the uploaded works or other subject-matter, irrespective of the means used therefor, and therefore act in an active way. As a consequence, they cannot benefit from the liability exemption provided for in Article 14 of Directive 2000/31/EC. The definition of online content sharing service providers under this Directive does not cover microenterprises and small sized enterprises within the meaning of Title I of the Annex to Commission Recommendation 2003/361/EC and service providers that act in a non- commercial purpose capacity such as online encyclopaedia, and providers of other subject-matter are used as well as their possibilities to get an appropriate remuneration for it, since some user uploaded content services do not enter into licensing agreements on the basis that they claim to be covered by the 'safe-harbour' exemption set out in Directive 2000/31/EC".

[26] Article 13, on the 'Use of protected content by online content sharing service providers storing and giving access to large amounts of works and other subject-matter uploaded by their users': "Without prejudice to Article 3(1) and (2) of Directive 2001/29/EC, online content sharing service providers perform an act of communication to the public. They shall therefore conclude fair and appropriate licensing agreements with right holders".

[27] See *Commission Staff Working Document Impact Assessment on the Modernisation of EU Copyright Rules*, cit., 147: "The above obligations will be without prejudice to liability regimes applicable to copyright infringements and the application of Article 14 ECD. In particular, with regard to services that are covered by Article 14 ECD, the obligation to put in place content identification technologies would not take away the safe harbour provided that the conditions of Article 14 are fulfilled. The notice and takedown regime will continue to apply for hosting service providers covered by Article 14 with respect to content not covered by agreements or in cases where the content is not properly identified". For a first opinion on the matter see [61].

[28] The spread of the social networks and the advent of other types of platforms that help the users in their search for websites, goods or other services – the quality of which is dependent on their collaborative nature – must be linked to a phenomenon of concentration of market power (which, in short, distinguishes every 'cycle' in the history of mass communication means), posing a series of problems of political and regulatory nature: on this matter see [62]; [63]; [64]; [65]; [66]; [67]; [68].

[29] See e.g. Italian Antitrust Authority, 12 May 2017, cases no. PS10601 and CV154, WhatsApp. *Il Diritto dell'informazione e dell'informatica*, 2, 2017, 390; KG Berlin, 8 April 2016, in *Multimedia und Recht*, 2016, 601.

[30] See the *Proposal for a Directive of the European Parliament and of the Council on certain aspects concerning contracts for the supply of digital content*, COM(2015) 634 final. On this matter see [69]; [70]; [71]; [72]; [73].

[31] Should the aforementioned draft law enter into force, it appears that would in any case be applied in a subsidiary or supplementary way, always within the respect for the general principles for the protection of personal data, such as those regarding the revocability of the consent of the data subject: see e.g. Recital no. 11, Directive 2011/83/UE on consumer rights: "This Directive should be without prejudice to Union provisions relating to specific sectors, such as medicinal products for human use, medical devices, privacy and electronic

communications, patients' rights in crossborder healthcare, food labelling and the internal market for electricity and natural gas".

[32] "The injured party can demand specific redress when this is wholly or partially possible. The court, however, can order that the redress be made only by providing an equivalent, if specific redress would prove to be excessively onerous for the debtor". On the matter see [74]; [75]; [76]; [77].

[33] On the goals of effective prevention connected to Tort Law system see [78].

[34] Frosini [79], p. 111, considers a robot endowed with artificial intelligence as a moral subject, in conflict between inner consciousness (characteristic of man born from ζωή) and outer consciousness (technical upgrading of the first). See also [80]; [81]; [82]; [83].

[35] Following this reasoning, we could add also a different definition of the cooperative approach to this inclusive sense of digital solidarity, stemming from the right to access and communicate online. In other words, to the concept of 'computer freedom' identified first by Vittorio Frosini [84], we can add the notion of 'online' or 'digital' solidarity, by which the transit of legal information from one area of the Internet to another leads to an increase in the overall value of the system and therefore of its functioning and the beneficial effects that it can have on society. On the concept of solidarity in the changed technological and social model, see [85], p. 115, in which this inclusive nature is highlighted in the context of the advent of new relational goods and the redistribution of power; in argument see also [86].

References

[1] Branscomb, A. W. (1995). Anonymity, Autonomy, and Accountability: Challenges to the First Amendment in Cyberspaces. *The Yale Law Journal, 104*(7), 1639-1679.

[2] Nicoll, C., Prins, J. E. J. & van Dellen, M. J. M. (Eds.) (2003). *Digital Anonymity and the Law*. TMC Asser Press, xiv+307.

[3] Posner, R. A. (2008). Privacy, Surveillance, and Law. *University of Chicago Law Review, 75*, 245.

[4] Krausova, A. (2008). Identification in Cyberspace. *Masaryk University Journal of Law and Technology, 2*, 83.

[5] Choi, B. H. (2012). The anonymous internet. *Maryland Law Review, 72*, 501.

[6] Zeno-Zencovich, V. (2014). Anonymous Speech on the Internet. In Koltay, A. (Ed.). *Media Freedom and Regulation in the New Media World*, 103-116.

[7] Finocchiaro, G. (2010). Anonimato. *Digesto delle Discipline Privatistiche–Sezione Civile. Aggiornamento*. UTET, 12-20.

[8] Codiglione, G. G. (2013). Indirizzo IP, Reti Wi-Fi e responsabilità per illeciti commessi da terzi. *Il Diritto dell'informazione e dell'informatica, 29*(1), 107-143.

[9] Resta, G. (2014). Anonimato, responsabilità, identificazione: prospettive di diritto comparato. *Il Diritto dell'informazione e dell'informatica, 30*(2), 171-205.

[10] Floridi, L. (2014). *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press.

[11] Baudrillard, J. (1968). *Les système des objects*. Denoël-Gonthier.

[12] Ellul, J. (1977). *Le Système technicien*. Calmann-Lévy, 138.

[13] Ferraris, M. (2014). *Documentalità: perché è necessario lasciar tracce*. Laterza.

[14] Lessig, L. (2006). *Code: version 2.0*. New York.

[15] Glenn, H. P. (2014). *Legal Traditions of the World: Sustainable Diversity in Law*. Oxford University Press.

[16] Wiener, N. (1954). *The Human Use of Human Beings: Cybernetics and Society*. Anchor.

[17] Licklider, J. C. R. (1963). *Memorandum for Members and Affiliates of the Intergalactic Computer Network*, April 23.

[18] Baran, P. (1964). On Distributed Communications Networks. *IEEE Transactions on Communications Systems, 12*(1), 1-9.

[19] Saltzer, J. H., Reed, D. P. & Clark, D. D. (1984). End-to-end Arguments in System Design. *Technology, 100*, 0661.

[20] Nef, J. U. (1958). *Cultural Foundations of Industrial Civilization*. CUP Archive.

[21] Ellul, J. (1954). *La technique ou l'enjeu du siècle*. Armand Colin.

[22] Irti, N. (2013). *L'uso giuridico della natura*. Laterza.

[23] Irti, N. (2014). *Nichilismo giuridico*. Laterza.

[24] Rodotà, S. (2018). *Vivere la democrazia*, Laterza, 17.

[25] Samuelson, P. A. (1954). The Pure Theory of Public Expenditure. *The Review of Economics and Statistics*, 387-389.

[26] Arrow, K. J. (1962). Economic Welfare and The Allocation of Resources for Invention. *The Rate and Direction of Inventive Activity*. Princeton University Press and NBER.

[27] Stiglitz, J. E. (2002). Information and the Change in the Paradigm in Economics. *American Economic Review, 92*(3), 460-501.

[28] Zeno-Zencovich, V. (1993). Informazione (profili civilistici). *Digesto delle Discipline Privatistiche*. UTET, 420.

[29] Pardolesi, R. & Motti, C. (1991). L'informazione come bene. In De Nova, G., Inzitari, B., Tremonti, G. & Visentini, G. (Eds.). *Dalle res alle new properties*. Giuffrè, 37.

[30] Giannone Codiglione, G. (2015). Libertà d'impresa, concorrenza e neutralità della rete nel mercato transnazionale dei dati personali. *Il Diritto dell'informazione e dell'informatica*, 905.

[31] Zeno-Zencovich, V. & Giannone Codiglione, G. (2016). Ten Legal Perspectives on the "Big Data Revolution". *Concorrenza e mercato, 23*, 29-57.

[32] Lipari, N. (2013). *Le categorie del diritto civile*. Giuffrè, 125.

[33] Calabresi, G. & Melamed, A. D. (1972). Property Rules, Liability Rules, and Inalienability: One View of the Cathedral. *Harvard Law Review, 85*, 1089-1128.

[34] Sica, S. & Giannone Codiglione, G. (Eds.) (2018). *Security and Hate Speech. Personal Safety and Data Security in the Age of Social Media*. Il Mulino.

[35] Sartor, G. (2017). *Providers Liability: From the eCommerce Directive to the Future*.

[36] Resta, G. (2007). Identità personale ed identità digitale. *Il Diritto dell'informazione e dell'informatica, 3*, 511 ff.

[37] Di Majo, A. (1999). Il trattamento dei dati personali tra diritto sostanziale e modelli di tutela. In Cuffaro, V., Ricciuto, V. & Zeno-Zencovich, V. (Eds.). *Trattamento dei dati e tutela della persona*. Giuffrè, 225.

[38] Sica, S., D'Antonio, V. & Riccio, G. M. (Eds.) (2016). *La nuova disciplina europea della privacy*. Cedam, 55.

[39] Ghidini, G. (2018). *Rethinking Intellectual Property: Balancing Conflicts of Interest in the Constitutional Paradigm*. Edward Elgar Publishing.

[40] Moscati, L. (2017). Sfide tecnologiche e diritto d'autore tra riferimenti storici e direttive europee. *Rivista italiana di scienze giuridiche*, 443.

[41] Giannone Codiglione, G. (2017). *Opere dell'ingegno e modelli di tutela. Regole proprietarie e soluzioni convenzionali*. Giappichelli.

[42] Giannone Codiglione, G. (2017). I motori di ricerca. *Annali italiani del diritto d'autore, della cultura e dello spettacolo*, 395.

[43] Riccio, G. M. & Giannone Codiglione, G. (2018). Ancillary Copyright and Liability of Intermediaries in the EU Directive Proposal on Copyright. *Comparazione e diritto civile*.

[44] Yen, A. C. (1999). Internet Service Provider Liability for Subscriber Copyright Infringement, Enterprise Libility, and the First Amendment. *Georgetown Law Journal, 88*, 1833.

[45] Scruers, M. (2002). The History and Economics of ISP Liability for Third Party Content. *Virginia Law Review, 88*, 205.

[46] Landes, W. & Lichtman, D. (2003). Indirect Liability for Copyright Infringement: An Economic Perspective. *J.M. Olin Law & Economics Working Paper no. 179 (2nd series)*. Chicago.

[47] Riccio, G. M. (2002). *La responsabilità civile degli internet providers*. Giappichelli.

[48] Lanier, J. (2014). *Who Owns the Future?* Simon and Schuster.

[49] Mezzanotte, F. (2017). Access to Data: The Role of Consent and the Licensing Scheme. In Lohsse, S., Schulze, R. & Staudenmayer, D. (Eds.). *Trading Data in the Digital Economy: Legal Concepts and Tools*. Nomos, 159-187.

[50] Bergé, J. S., Grumbach, S. & Zeno-Zencovich, V. (2018). The 'Datasphere', Data Flows beyond Control, and the Challenges for Law and Governance. *European Journal of Comparative Law and Governance, 5*(2), 144-178.

[51] Zeno-Zencovich, V. (1993). Identità personale. *Digesto delle Discipline Privatistiche*. UTET, 294.

[52] Finocchiaro, G. (2010). Identità personale (diritto alla). *Digesto delle Discipline Privatistiche*. UTET, 721.

[53] Resta, G. (2011). The New Frontiers of Personality Rights and the Problem of Commodification: European and Comparative Perspectives. *Tulane European and Civil Law Forum, 26*, 33.

[54] Calabresi, G. (1970). *The Costs of Accidents: A Legal and Economic Analysis*. Yale University Press.

[55] Ponzanelli, G. (1992). *La responsabilità civile. Profili di diritto comparato*. Il Mulino.

[56] Peifer, K. N. (2016). Beseitigungsansprüche im digitalen Äußerungsrecht-Ausweitung der Pflichten des Erstverbreiters. *Neue Juristische Wochenschrift*, 23-25.

[57] Spindler, G. (2012). Persönlichkeitsschutz im Internet-Anforderungen und Grenzen einer Regulierung. In Gutachten, F. (Ed.). *Ständige deputation des Deutschen Juristentages Verhandlungen des 69*. Beck.

[58] Hartmann, A. (2009). *Unterlassungsansprüche im Internet: Störerhaftung für nutzergenerierte Inhalte*. Beck, Vol. 75.

[59] Neuhaus, S. (2011). *Sekundäre Haftung im Lauterkeits-und Immaterialgüterrecht: dogmatische Grundlagen und Leitlinien zur Ermittlung von Prüfungspflichten*. Mohr Siebeck, Vol. 50.

[60] Leistner, M. (2012). Common Principles of Secondary Liability? *Common Principles of European Intellectual Property Law*, 117-146.

[61] Angelopoulos, C. (2017). *On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2947800.

[62] Wu, T. (2010). *The Master Switch: The Rise and Fall of Information Empires*. Vintage.

[63] Gorz, A. (2003). *L'immatériel: connaissance, valeur et capital*. Ed. Galilée.

[64] Cairncross, F. (1997). *The Death of Distance: How the Communications Evolution Will Change Our Lives*. Harvard Business School Press.

[65] Ryan, J. (2010). *A History of the Internet and the Digital Future*. Reaktion Books.

[66] Rifkin, J. (2014). *The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism*. Macmillan.

[67] Zeno-Zencovich, V. & Mezzanotte, F. (2008). Le reti della conoscenza: dall'economia al diritto. *Il Diritto dell'informazione e dell'informatica*, 2, 141.

[68] De Franceschi, A. & Lehmann, M. (2015). Data As Tradeable Commodity and New Measures for Their Protection. *Italian Law Journal*, 1, 51.

[69] Resta, G. & Zeno-Zencovich, V. (2018). Volontà e consenso nella fruizione dei servizi in rete. *Rivista trimestrale di diritto e procedura civile*, 2, 411.

[70] Langhanke, C. & Schmidt-Kessel, M. (2015). Consumer Data as Consideration. *Journal of European Consumer and Market Law, 4*(6), 218-223.

[71] Twigg-Flesner, C. (2016). Disruptive Technology – Disruptive Law? In De Franceschi, A. (Ed.). *European Contract Law and the Digital Single Market*. Intersentia, 21, 40.

[72] Metzger, A. (2017). Data As Counter-Performance: What Rights and Duties for Parties Have. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 8*, 2.

[73] Zeno-Zencovich, V. (1993). Profili negoziali degli attributi della personalità. *Il Diritto dell'informazione e dell'informatica*, 3, 545.

[74] Salvi, C. (2005). *La responsabilità civile*. Giuffrè, 250.

[75] Di Majo, A. (2003). *Problemi e metodi del diritto civile. 3. La tutela civile dei diritti*. Giuffrè, 261.

[76] Marella, M. R. (2000). *La riparazione del danno in forma specifica*. Cedam.

[77] Stoll, H. (1972). Consequences of Liability: Remedies. In Tunc, A. (Ed.). *International Encyclopedia of Comparative Law*, Vol. XI.

[78] Monateri, P. G. (2013). La natura del 'politico' e il 'problema' della responsabilità civile. In Alpa, G. & Roppo, V. (Eds.). *La vocazione civile del giurista: saggi dedicati a Stefano Rodotà*. Laterza, 243.

[79] Frosini, V. (1968). *Cibernetica diritto e società*. Edizioni di comunità, 111.

[80] Solum, L. B. (1991). Legal Personhood for Artificial Intelligences. *The North Carolina Law Review, 70*, 1231.

[81] Teubner, G. (2007). *Rights of Non-humans? Electronic Agents and Animals As New Actors in Politics and Law*. Max Weber Lecture No. 4.

[82] Rodotà, S. (2012). *Il diritto di avere diritti*. Laterza, 312, 341.

[83] Sartor, G. (2009). Cognitive Automata and the Law: Electronic Contracting and the Intentionality of Software Agents. *Artificial Intelligence and Law, 17*(4), 253.

[84] Frosini, V. (1984). 1984: L'informatica nella società contemporanea. *Informatica e diritto*, 3, 7-15.

[85] Rodotà, S. (2014). *Solidarietà*. Laterza, 115.

[86] Sunstein, C. R. & Ullmann-Margalit, E. (2001). Solidarity Goods. *Journal of Political Philosophy, 9*(2), 129-149.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190006

Closing the Awareness Gap Between IT Practice and IT Law

François MESTRE [a], Victor RODRÍGUEZ-DONCEL [b] and Pompeu CASANOVAS [c,a]
[a] *Unívsersitat Autónoma de Barcelona (Spain)*
[b] *Universidad Politécnica de Madrid (Spain)*
[c] *School of Law, La Trobe University (Australia)*

Abstract. Some of the ordinary activities of IT practitioners require a certain degree of knowledge of IT law. Assuming these professionals will acquire legal knowledge better if expressed in terms familiar to them, this Chapter explores different manners of organising and presenting legal knowledge for its better cognition by IT professionals. This proposal features data models and knowledge organisation rooted in the specific legal theory of critical legal positivism of Kaarlo Tuori. It has been evaluated with an experiment, where BSc students in Computer Science have been provided with models and reference material describing the EU legislation on cookies, and have been asked specific questions. In sight of the new theoretical framework and the experiment results, we postulate that models and ontologies can bridge the knowledge gap and serve as lingua franca between the legal and the IT profession.

Keywords. legal knowledge, critical legal positivism, cognition of law, semi-formal models

1. Introduction

A model is a representation of a reality, an abstraction, a simplification, a depiction. Modelling law can be a necessity for a number of reasons: legal drafting, analysis of court cases, development of computer programs implementing or enforcing law, teaching law [1]. This Chapter discusses the pros and cons of using formal or semi-formal models for representing the law with didactic purposes, specifically when introduced to IT practitioners.

The use of ontologies and semi-formal models in education environments is not new [2]. Semantic Web Technologies have been used for e-Learning [3], to align curriculum and syllabuses with learning objectives [4] or to support adaptive learning [5]. This work proposes their novel application to the legal domain – which has its own idiosincracy – and addressing a specific collective – IT experts. The contribution of this Chapter revolves around the two proposed ideas: (i) that this target group is likely to understand better UML diagrams and related documentation and (ii) that legal knowledge representation must consider a theoretical framework and lean on a rich tradition: languages and methods for software and ontological engineering make explicit an idea that has been traditionally part of legal thinking, the idea that legal concepts have a structure and are linked one to another [6].

We engage Tuori's critical legal positivist theory [7], which conceives law as a multi-layered phenomenon, as a feasible way to pre-conceptual modelling[1]. Tuori refers to an 'upper level' concerned with legislative acts and case law, a "middle level, mediating level in the law" related to the practical knowledge which lawyers and judges require to interpret the law, and finally at the 'lower level', or the 'deep structure' of law, a compendium of the most basic principles and habits of mind by which we think and argue about the law. Critical legal positivism acknowledges the two faces of the law: "on the one hand, the law is a symbolic normative phenomenon, and on the other hand, it can be defined as a set of specific social practices". This Chapter assumes that one interesting mapping from a pragmatic point of view is that of the mediating legal culture level, where concepts have a cultural existence even for laymen.

Some of these legal concepts must be apprehended by IT professionals who need to comply with the applicable legislation in their daily activity, but do not have the time or the background to learn about law[2]. We contend that the intermediation of semi-formal models and their derived diagrams and schemas are helpful towards improving the legal literacy of IT professionals, contributing to decrease the number of unintended law breaches.

As a case study, this work considers the legal regime on cookies as sample of the domain of IT law. Models and associated reference material meant to be employed by human users to learn this subject will be designed and presented. We pretend to assess the effectiveness of ontologies and Unified Modelling Language[3] (UML) as representations of IT law to cover the legal knowledge needs of IT practitioners. Are UML models and derived documentation suitable to teach the domain of IT law to an audience of non-legal experts? In order to answer this question, this Chapter is organised as follows. Section 2 frames the work first, describing the theory of critical legal positivism. Section 3 describes the methodology followed, both for creating the models and for evaluating their didactic aptitude. Section 4 describes the metamodel created for this project in the context of the state of the art of legal ontologies. Finally Section 5 presents the UML model of the EU legal regime on cookies resulting from this project and the experiment designed for its evaluation.

2. Framework: Critical Legal Positivism

Epistemology is a branch of philosophy that deals with the questions of "How do we know that we know?" Because our endeavour aims at providing legal knowledge to non-legal experts, we can ask from an epistemological perspective how do we know that non-legal experts understand what is a right, an obligation, a norm, a doctrine and all the concepts implicit in law that lawyers learn the hard way through their legal education.

[1] Tuori's contribution tries to reconcile the public civil law approach with the realist socio-legal one, the rule of law with the European *Rechtsstatt*, cfr. [8]. Tuori [9] has pointed out his starting critical stance: "[...] how can the law limit and discipline the exercise of state power, if the law itself emanates from this very state power, as it according to the dominant positivist understanding does?".

[2] Sometimes they are not even fully aware, as demonstrated by the Deloitte 2016-2017 CIO survey, where 'Governance and compliance' was only chosen in 6th place by CIOs as one of the organisational capabilities for success.

[3] http://www.omg.org/spec/UML/2.5/.

As posited by Valente and Breuker [10], “the perspectives and conceptualizations used in Legal Theory have the advantage of having been discussed and debugged in the course of years, for a research community whose work is centered on these problems. For the AI & Law community to create or use ontologies without regard to Legal Theory is a certain path to reinvent the wheel”. In their article, the authors propose a review of the ontological views about the legal order from the perspectives of the main legal theories. Legal positivism describes the legal order as characterized by its formal normative sources, as opposed to the legal realism which focuses instead on “what happens in Court”. A third avenue would be focusing on the legal discourse.

Following the dominant positivist approach, the authors ([10], p. 143) take as reference the works of Hart, Kelsen, Bentham or Hohfeld and their respective theories of norms; “because legal knowledge is closely associated to the formal sources of law (statutes, jurisprudence, etc.), ontologies of law may adopt (and frequently do so) as a phenomena not the legal phenomena in legal practice but these sources”. Other work on legal theory, sources of law and semantic Web, by Alexander Boer [11], has also approached the legal order via the intermediation of taxonomies of legal norms, distinguishing between legislative acts, institutional rules, constitutive rules and normative rules. We subscribe to the position of Boer who concludes, his PhD thesis ([11], p. 271) with the acknowledgement that sources of law are only “indirect, incomplete and approximate representation of the normative order”, and discounts the direct relation (or isomorphism) between formal normative sources and the legal order. This comprehensive approach is shared by cognitive [12] and socio-legal [13] approaches to legal ontologies. Rinke Hoekstra has shown both the need and difficulty of knowledge representation and design patters in ontology reengineering [14].

The domain of IT law has a number of characteristics against a conception of such an isomorphism between the sources of law and the legal order. Indeed, a dualistic separation between the world of ‘is’ and the world of ‘ought’ would fall short of explaining practices which are typical of the domain of IT law and which blur this strict ontological duality. We are referring for example to the practice of self-regulation, which is widespread in the IT sector, the interplay between technology and law so relevant for the question of regulability [15] or the atypical involvement of online intermediaries in enforcement as regulatory gatekeepers [16].

The domain of IT law is also characterized by its very low contentious nature and scarce volume of judiciary decisions – in our case, there are very few cases reported on cookies[4]. As a matter of fact, the level of contentious resolutions of conflicts is very low in the IT domain compared to other branches of law, and parties usually prefer to turn to other conflict resolution methods such as more discrete and predictable Alternative Dispute Resolution solutions [17].

Parting from the positivist normative definition of Law, our work preliminary employs the conception of the *legal order* proposed by Kaarlo Tuori, which combines positivism and legal practice and takes the very structure of the legal order as the defining criteria of law. Tuori’s critical positivist theory conceives law’s structure as partitioned into three clearly defined layers [7]. The *surface* layer consists in the symbolic normative order made of rules and principles. The *intermediate layer* refers to the legal culture as the practical legal knowledge (or practical consciousness) that legal experts (e.g.,

[4] An example is represented by the 2017 Annual Report of the French Data Protection Authority.

lawyers, among other legal professions) acquire through education and practice and require to participate in legal community. Finally, the *deep structure of law* also refers to practical knowledge, even if usually subconscious to legal actors. Figure 1 summarizes the perspective of the legal order that will serve as theoretical framework for our models of IT law.

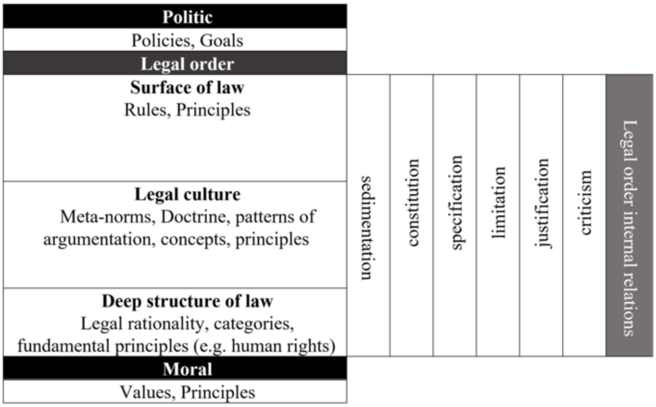


Figure 1. Critical legal positivism epistemological framework

3. Methodology

3.1. Methodology and Toolset for Knowledge Representation

The enterprise at hand is one of knowledge representation, a description of the objects, concepts, rules and relations that exist in the domain of IT law. Because it is impossible to represent law in all of its details, this implies a level of conceptualization and abstraction. For that purpose, the discipline of knowledge engineering has borrowed from philosophy the notion of ontology, which in Philosophy means an account of Existence. Knowledge engineers create computer ontologies to specify conceptualizations of specific domains [18]. From this point of view, there are some differences between philosophical and computational ontologies: “we refer to an ontology as a special kind of information object or computational artifact (...) the account of existence in this case is a pragmatic one: For AI systems, what ‘exists’ is that which can be represented. Computational ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes” [19].

Soon after computer ontologies were developed first in the early 90’s, routine procedures for their definition were given. The ontologist’s work was systematically analyzed, and the new discipline of Ontology Engineering was born.

From all the available methodologies for ontology design [20], the NeOn Methodology has been selected for its comprehensive coverage of the entire ontology development lifecycle, the systematization of requirement specifications activities, the availability of documentation and detailed guidelines [21].

The NeOn Methodology organizes the design work in a waterfall-like manner [22]. Applying the NeOn methodology for the development of a model of the EU legal regime on cookies, a requirement document was developed spelling out the purpose, intended audience and use of the model. This document also identifies the definitions that the model should provide and the competency questions that it should answer. A summary of these requirements is provided in Table 1.

Table 1. Requirements specifications for the model on the EU legal regime on cookies

Purpose	Inform online providers on legal requirements they should consider before placing cookies on the terminal equipment of Internet users in the European Union (EU legal regime on cookies)
Scope	The model uses a general granularity at the level of legal provisions, representing rights, obligations, prohibitions, exceptions, constraints, enforcement procedures, further interpretation, requisites, legal sources.
Intended end-users	Webmasters, CIOs and in-house counsellors for online service providers Advertising network providers and publishers Regulators Member States Data Protection Authorities
Intended uses	Training Sharing of consensual knowledge on the EU legal regime on cookies Support to legal opinion on compliance with EU legislation Legal theory
Competency questions	To which technologies is Art. 5(3) of the E-Privacy Directive (2002/58/EC) applicable? What are the legal grounds to store cookies on the terminal equipment of a user? What are the legitimate purposes of such devices? What are the illegitimate purposes of such devices? What are the rights of the users of the terminal equipment? What are the elements of a valid prior informed consent? Who has the obligation to seek prior informed consent? What are the obligations of the advertising network providers and publishers? What are the modalities of the prior informed consent? Can the consent serve for several cookies? For how long does the consent serve? What mechanisms can be used to seek consent? Can valid consent be implicit (e.g. obtained through browser settings)? When can cookie be exempt from the principle of informed consent? For what purposes can (or can not) the exemption be applied? What is applicable in these exceptional cases? What are the possible enforcement action in case of non compliance?
Defined terms (definitions not provided here)	Cookies (technical definition) Cookies and similar tracking technologies (or related technologies) (legal definition) Characteristics of cookies Online Behavioural Advertising Terminal equipment Setting cookies (syn. planting cookies) Third party Web site operator Users

Representing knowledge with models requires to select a formalism from the available formal and semi-formal alternatives. This project considered formalisms and methods developed for software and ontological engineering. The industry standard for software engineering is the semi-formal UML, whereas computer ontologies rely on the formal Ontology Web Language (OWL). The suitability for this project of these modelling languages was assessed against a set of common requirements.

These requirements not only considered the expressivity of the language itself (UML/OWL) but also the capabilities of a toolbox of reference (Protégé vs Sparx Enterprise Architect). The requirements were divided into three domains related to the ability to *model* (e.g. ability to work with modules), *share* (e.g. quality of automatic documentation) and maintain (e.g. *manage versions*) the resulting modelling artifacts. Under this pragmatic perspective, it was concluded that the semi-formal UML was the most adequate formalism for the needs of this study.

3.2. Methodology for evaluation of the pedagogic efficiency

A rigorous pedagogic methodology is needed to assess whether this project achieves its objective, which is to provide models and reference material that can actually teach law to IT practitioners. Following the academic best practice methodology of constructive alignment in higher Education advanced by Biggs [23], it is essential to refer to taxonomies of learning outcomes, such as the SOLO taxonomy [24], Hailikari's levels of understanding or Bloom's taxonomy of educational objectives [25]; [26]. In what follows, we will refer to the revised version of Bloom's taxonomy of educational objectives [27] which explains that human learning follows different levels of increasing complexity: *Remembering*, *Understanding*, *Applying*, *Analyzing*, *Evaluating* and *Creating*. Students need to reach one level before being able to access the next one.

In what follows, we will justify the assumptions that leads us to limit the learning outcomes for this project to the bottom three levels of the Bloom's taxonomy, *Remembering*, *Understanding* and *Applying*.

To that end, we need to make an assumption about the learning preference of IT practitioners based on the empirical research performed by Joe Peppard [28] who has benchmarked the personality traits of hundreds of Chief Information Officers (CIO) against the Myers-Briggs Types MBTI®. The result of Peppard's experiment is that 70% of CIOs are of profile type ISTJ (compared to 12% ISTJ in the rest of the population).

Having such a distinctive personality trait of CIOs is helpful to determine their learning preferences. D. Dunning has established a direct correlation between MBTI® types and their learning preference [29]. In order to learn efficiently, people of profile type ISTJ need [30]:

- Clarity and structure;
- Detailed and concrete information;
- Clear objectives and expectations (e.g. learning outcomes);
- Self-study with sufficient time to absorb the material on their own before engaging in group activities;
- Practical application of the knowledge learned (e.g. examples, case studies).

Based on this last point, we conclude that *Applying* should be the highest level for our learning outcomes in order to satisfy the dominant learning preference of IT practitioners, whereas the three topmost levels of the Bloom's taxonomy are excluded from the learning outcomes of this project, as they are clearly meant to develop the legal culture of the aspiring lawyers and exceed the legal knowledge needs of the IT practitioner in non-litigious matters.

In the following Section, we will review the reusable elements available from the state of the art in the discipline of legal knowledge engineering, and how these elements converge, together with the theoretical framework described in Section 2, to constitute "Metamodel for an ontology of IT law", as described in the Section 4.2.

4. Data models and related documentation

4.1. Legal ontologies

The seminal definition of computer ontologies is often attributed to Gruber who described computer ontologies as "an explicit specification of a conceptualization" [18]. However, the notion of ontology is polymorph and there is a wide variety of possible definitions to choose from, as shows the recompilation proposed by Nuria Casellas [20] in her thesis on legal ontologies engineering ([20], p. 57). She ascribes to the working definition by Uschold and Jasper (1999) [31]: "An ontology may take a variety of forms, but necessarily it will include a shared vocabulary of terms, and some specification of their meaning. This includes definitions and an indication of how concepts are interrelated which collectively impose a structure on the domain and constrain the possible interpretations of terms".

By legal ontology, we mean in this article ontologies that have been developed to represent legal knowledge for a variety of purposes such as search and retrieval of information, reasoning applications, or communication between people or organizations. The development of legal ontologies has been a prolific area of research, resulting in growing number of available ontologies which can be inventoried following the specific typology proposed by Nuria Casellas. According to her conclusions, most legal ontologies are domain ontologies and focus on particular applications based on domain knowledge, aiming at representing knowledge towards a specific application. The observed uses of legal ontologies include information search and retrieval (e.g. semantics), reasoning, communication between systems and organizations (e.g. interoperability), with limited occurrence of legal ontologies for communication between humans for example for a use case such as education. Some ontologies have been developed at the core level (e.g. LKIF Core) aiming mainly at knowledge reuse with more theoretical legal knowledge in mind, but reuse has not been the main feature in the development of legal ontologies.

Legal ontologies have been recently evolving towards a manageable and useful way of implementing regulatory systems [32], with several examples coming from several EU projects[5], managing legal queries and services on the EurLex platform CELLAR [33]; [34], and the implementation of rights contained into new EU important regulations (such as the GDPR protections) [35].

Following the adopted knowledge engineering methodology (NeON), one crucial step consists in identifying sources of knowledge, either in formal or semi-formal languages, reusable for the purpose of the design. Building on the epistemological framework introduced in Section 2, this multidisciplinary study shall focus its epistemological viewpoint into each of the different layers of law, the surface, the legal culture and the deep structure of law, as well as from other disciplines tied to law such as politics. Each of these layers and disciplines constitutes a distinguished medium characterised by its

[5] Such as RESPECT, OPENLaws, and LYNX.

own rationality and functional language, its own ontology. The analysis of the state of the art of the existing legal ontologies shall provide us with the shortlist of candidate ontologies for reuse for this project.

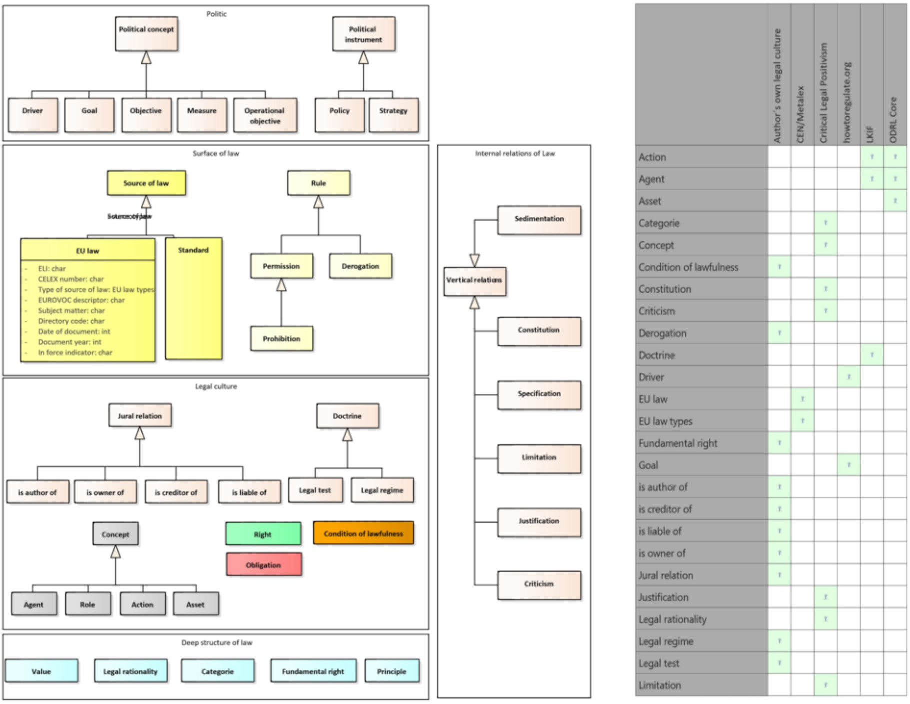
In Table 2, we list some of the relevant ontologies for each of these layers and disciplines.

Table 2. Different types of legal ontologies according to Tuori's layers

Layer / Discipline	**Sources of knowledge**
Politics	Regulatory Institute (Manfred Kohler) handbook Howtoregulate.org
Surface of law	CEN/Metalex, FRBR, ELI, ECLI, ODRL vocabulary, LegalRuleML
Legal culture and Deep structure of law	Policy language expression: ODRL core (concepts) LKIF [Deliverable 1.4 OWL Ontology of Basic Legal Concepts (LKIF-Core)]

4.2. Metamodel for an ontology of IT law

Epistemological choices introduced in Section 2 and the elements reused from the ontologies described above converge into a metamodel for our IT law (Figure 2).



	Author´s own legal culture	CEN/Metalex	Critical Legal Positivism	howtoregulate.org	LKIF	ODRL Core
Action					†	†
Agent					†	†
Asset						†
Categorie			†			
Concept			†			
Condition of lawfulness	†					
Constitution			†			
Criticism			†			
Derogation	†					
Doctrine					†	
Driver				†		
EU law		†				
EU law types		†				
Fundamental right	†					
Goal				†		
is author of	†					
is creditor of	†					
is liable of	†					
is owner of	†					
Jural relation	†					
Justification			†			
Legal rationality			†			
Legal regime	†					
Legal test	†					
Limitation			†			

Figure 2. Metamodel for an ontology of IT law with indication of provenance of ontological commitments

5. Evaluation

In the seminal work of Gruber [18] in 1993 on formal ontologies, he determined that their primary role was to “support knowledge sharing activities”. Even if a number of other use cases has justified the use of ontologies since the time of Gruber, ontologies are still primarily used to share the common understanding of a domain among people or software agents.

However, not many specific uses of computer ontologies have been reported where ontologies have mediated between humans with a different background knowledge, nor computer ontologies have been much used for teaching and learning [36].

This study experiments with the use of models representing IT law for the purpose of bridging the legal knowledge needs of IT practitioners. It takes as a test bed the EU legislation on cookies (ePrivacy) of which it provides an ontology interconnecting terms, concepts, categories and principles of IT law. The outcome is a self-learning material consisting of the UML model accessible via an HTML user interface[6].

The ontology has been formally specified using UML notation (semi-formal language), see Figure 3, whereas an early version of the model was specified in the formal language OWL[7].

5.1. Experimental setting

An experiment was run with a class of engineers from the Universidad Politécnica de Madrid in their last year of Master Degree. A summative assessment was used to assess the achievement of the learning outcomes.

The twenty-four students were divided into 3 groups:

- Group 1 (11 students): Students with access to the model of the EU legal regime on cookies as only source of knowledge. No time granted to familiarize with the model before answering the questions.
- Group 2 (7 students): Students with access to Google as only source of knowledge.
- Group 3 (6 students): Students without access to any source of knowledge.

The assessment method consisted in a quantitative assessment of the relative accuracy of answers of each student groups.

The students were asked to answer a quiz of nine questions[8] designed as formative assessment [37]. The total duration of the test was 20 minutes. None of the students had received recent courses on the topic of the EU legal regime on cookies.

In relation with the learning outcomes introduced in Section 3.1, the quiz was featuring questions one to four to assess the learning outcome of level 1 (*Remembering*), questions five to eight assessed the level 2 (*Understanding*) and question nine assessed the level 3 (*Applying*).

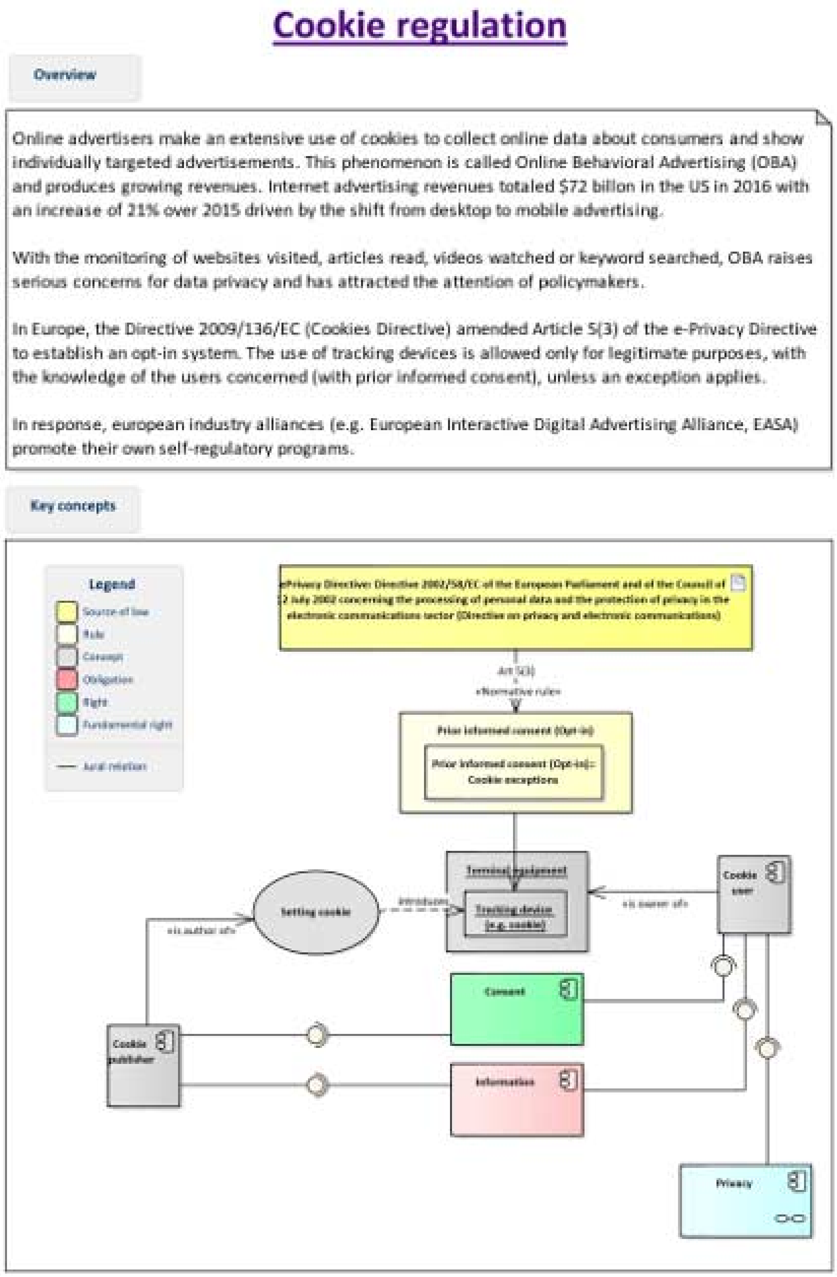
5.2. Results

The results of the assessment are provided in Figure 4.

[6] http://www.ciolaws.com/courses/cookies/ (username:cookies, password:1cookies1).

[7] http://www.ciolaws.com/ontologies/eulr-cookies.htm (25/04/2018).

[8] https://goo.gl/forms/qt0TpAsfT2gSJAQB3.



Cookie regulation

Overview

Online advertisers make an extensive use of cookies to collect online data about consumers and show individually targeted advertisements. This phenomenon is called Online Behavioral Advertising (OBA) and produces growing revenues. Internet advertising revenues totaled $72 billon in the US in 2016 with an increase of 21% over 2015 driven by the shift from desktop to mobile advertising.

With the monitoring of websites visited, articles read, videos watched or keyword searched, OBA raises serious concerns for data privacy and has attracted the attention of policymakers.

In Europe, the Directive 2009/136/EC (Cookies Directive) amended Article 5(3) of the e-Privacy Directive to establish an opt-in system. The use of tracking devices is allowed only for legitimate purposes, with the knowledge of the users concerned (with prior informed consent), unless an exception applies.

In response, european industry alliances (e.g. European Interactive Digital Advertising Alliance, EASA) promote their own self-regulatory programs.

Key concepts

Legend
Source of law
Rule
Concept
Obligation
Right
Fundamental right
Jural relation
ePrivacy Directive: Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)
Art 5(3)
«Normative rule»
Prior informed consent (Opt-in)
Prior informed consent (Opt-in):: Cookie exceptions
Terminal equipment
Tracking device (e.g. cookie)
Setting cookie
introduces
«is author of»
«is owner of»
Cookie user
Consent
Cookie publisher
Information
Privacy

Figure 3. Model of the EU legal regime on cookies (UML version)

Result of the assessment of the learning outcome level 1 (*Remembering*): questions from 1 to 4. Group 1 scored better than groups 2 and 3 on the questions from 1 to 4, which were questions of type ‘fill the gap’, which correct answer could be found in the model.

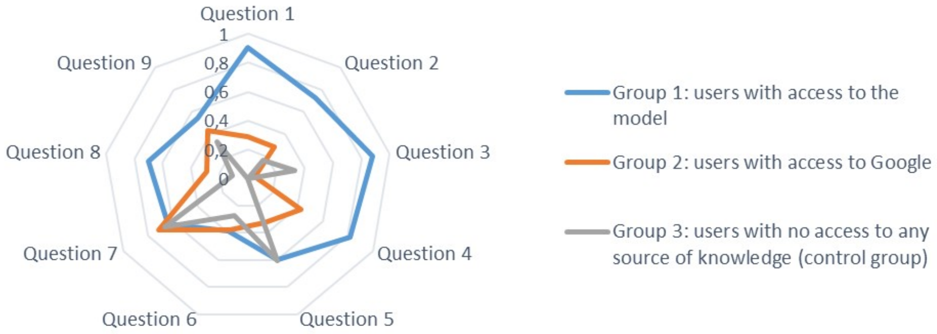


Figure 4. Results of quantitative assessment

Result of the assessment of the learning outcome level 2 (*Understanding*): questions from 5 to 7. Question 5 was essential to this test as it was asking "What are the right of the cookie users?" On this question, students from group 3 (without access to any source of knowledge) achieved identical scores as the students from group 1. This unexpected result allowed to identify a flow in the design in the representation of rights and obligations in the model. This aspect was subsequently improved in an updated version of the model. The three groups achieved similar results on questions 6 and 7. In these questions, the level of understanding of students in group 1 was challenged by using a vocabulary slightly distinct from the one used in the model, requiring them to paraphrase the question and recognize associations between similar concepts.

Analysis of learning outcome level 3 (*Applying*): question 9. The question 9 was testing the ability of applying the knowledge acquired on a practical case, by providing a short description of the EASA/IAB Code on cookies and asking if "The EASA/IAB Code provides users with consent options compliant with Article 5(3)". Group 1 scored better than groups 2 and 3, even though with little difference. Also, having only one single question to test this learning outcome, the experiment is inconclusive on that point and the results from this question need to be excluded from this experiment.

Overall Analysis: the model achieves to provide a good level of knowledge of the subject matter at level 1 (*Remembering*), as users are able to access knowledge in an autonomous manner, provided that the same language is used in the questions as in the model. The level of understanding of the students in Group 1 was challenged by testing the ability to paraphrase and relate similar concepts. However, the experiment was designed in a way that students had very limited time to read through the material provided, learn and understand its content. This explains that the students in Group 1 did not achieve higher scores than the other groups. The experiment cannot conclude on the achievement of the level 2 (*Understanding*) of the learning outcomes. For that purpose, a new experiment should be designed where students from Groups 1 and 2 would avail of identical amount of time for self-learning, prior to receiving the questions of the test.

6. Conclusion

IT executives today deal with more legal issues than ever before. These issues include compliance with privacy rules and data protection related measures, which are of capital importance for their daily professional activity. Complying with these legal requirements becomes a complex burden for IT executives who need to understand the legal objects, concepts, rules and their relations. Whereas the use of ontologies and IT-friendly diagrams for learning is not new, the legal knowledge organisation proposed in this Chapter has been grounded on legal theory to facilitate an accessible visual language able to offer an intuitive representation of IT law.

This work has required a multidisciplinary approach, with a legal standpoint relying on Tuori's critical positivist theory of law's levels systems, and a knowledge engineering standpoint borrowing languages and methods from software and ontological engineering to represent knowledge in an explicit, shareable and maintainable manner. The result of this effort has been materialized in a collection of data models and a specific experiment carried out with IT students. Its conceptual and experimental results are promising enough as to invite to further exploration.

Acknowledgements

This project has received funding from the project TIN2016-78011-C4-1-R, and partially from Meta-Rule of Law, DER2016-78108-P.

References

[1] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

[2] Casanovas, P. (2012). Legal Crowdsourcing and Relational Law: What the Semantic Web Can Do for Legal Education. *Journal of Australian Law Teachers Association, 5*(1-2), 159-176.

[3] Sampson, D. G., Lytras, M. D., Wagner, G. & Diaz, P. (2004). Ontologies and the Semantic Web for E-learning. *Educational Technology & Society, 7*(4), 26-28.

[4] Chung, H. S. & Kim, J. M. (2012). Learning Ontology Design for Supporting Adaptive Learning in e-Learning Environment. In *Proceedings of the International Conference on Information and Computer Networks (ICICN 2012), 27.* IACSIT Press, 148-152.

[5] Ahmed, S., Parsons, D. & Ryu, H. (2010). Supporting Adaptive Learning Interactions with Ontologies. In *Proceedings of the 11th International Conference of the NZ Chapter of the ACM Special Interest Group on Human-Computer Interaction*. ACM, 17-24.

[6] Sartor, G., Casanovas, P., Biasiotti, M. & Fernández-Barrera, M. (Eds.) (2010). *Approaches to Legal Ontologies: Theories, Domains, Methodologies*. Springer Science & Business Media, Vol. 1.

[7] Tuori, K. (2017). *Critical Legal Positivism*. Routledge, 192.

[8] Petrusson, U. & Glavå, M. (2008). Law in a Global Knowledge Economy - Following the Path of Scandinavian Sociolegal Theory. *Law and Society. Scandinavian Studies in Law, 53*, 94-133.

[9] Tuori, K. (2013). The Common Core of the Rule of Law and the Rechtsstaat. *Legal journal «Law of Ukraine»*, (4), 14-17.

[10] Valente, A. & Breuker, J. (1994). Ontologies: The Missing Link between Legal Theory and AI & Law. In H. Prakken, H., Muntjewerff, A. J. & Soeteman, A. (Eds.). *Legal Knowledge Based Systems: The Relation with Legal Theory. Proceedings of the 7th Jurix Conference*, Koninklijke Vermande, 138-150.

[11] Boer, A. (2009). Legal Theory, Sources of Law and the Semantic Web. In *Proceedings of the 2009 Conference on Legal Theory, Sources of Law and the Semantic Web*. IOS Press, 1-316.

[12] Breuker, J. & Hoekstra, R. (2011). A Cognitive Science Perspective on Legal Ontologies. In *Approaches to Legal Ontologies*. Springer, 69-81.

[13] Casanovas, P., Casellas, N. & Vallbé, J. J. (2011). Empirically Grounded Developments of Legal Ontologies: A Socio-legal Perspective. In *Approaches to Legal Ontologies*. Springer, 49-67.

[14] Hoekstra, R. (2010). The Knowledge Reengineering Bottleneck. *Semantic Web, 1*(1), 111-115.

[15] Lessig, L. (2006). *Code: version 2.0.* Basic Books.

[16] Rowland, D., Kohl, U. & Charlesworth, A. (2011). *Information Technology Law* (4th ed.). Routledge, 73.

[17] Koulu, R. (2016). *Dispute Resolution and Technology: Revisiting the Justification of Conflict Management*, https://helda.helsinki.fi/handle/10138/165460.

[18] Gruber, T. R. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* Knowledge Systems Laboratory, Stanford University. Technical Report.

[19] Guarino, N., Oberle, D. & Staab, S. (2009). What Is an Ontology? In *Handbook on Ontologies*. Springer, 1-17.

[20] Casellas, N. (2011). *Legal Ontology Engineering: Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge*. Springer Science & Business Media, *3*, 57.

[21] Suarez-Figueroa, M. C., Gomez-Perez, A., Motta, E. & Gangemi, A. (2012). Introduction: Ontology Engineering in a Networked World. In Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E. & Gangemi, A. (Eds.). *Ontology Engineering in a Networked World*. Springer, 1-6.

[22] Suárez-Figueroa, M. C., Gómez-Pérez, A. & Villazón-Terrazas, B. (2009). How to Write and Use the Ontology Requirements Specification Document. *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 966-982, http://link.springer.com/10.1007/978-3-642-05151-7_16.

[23] Biggs, J. & Tang, C. (2011). *Teaching for Quality Learning at the University*. Open University Press.

[24] Biggs, J. (2003). Aligning Teaching for Constructing Learning. *Higher Education Academy, 1*(4).

[25] Ferber, P. S. (2001). *Bloom's Taxonomy: A Structure for Structuring Legal Education*, https://papers.ssrn.com/abstract=271415.

[26] Gibson, M. T. (2012). A Critique of Best Practices in Legal Education: Five Things All Law Professors Should Know. *University of Baltimore Law Review, 42*(1).

[27] Anderson, L. W. & Krathwohl, D. R. (Eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). Longman.

[28] Peppard, J. (2014). *Why Chief Information Officers Can Struggle in a Leadership Role*. ESMT Knowledge, https://knowledge.esmt.org/article/why-chief-information-officers-can-struggle-leadership-role.

[29] Dunning, D. (2003). *Introduction to Type® and Learning*. CPP, 42.

[30] Humanmetrics (2019). *ISTJ Learning Style*, humanmetrics.com/personality/istj-learning-style.

[31] Uschold, M. & Jasper, R. (1999). A Framework for Understanding and Classifying Ontology Applications. In *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*, Vol. 2.

[32] Casanovas, P., Palmirani, M., Peroni, S., van Engers, T. & Vitali, F. (2016). Semantic Web for the Legal Domain: The Next Step. *Semantic Web, 7*(3), 213-227.

[33] Francesconi, E., Küster, M. W., Gratz, P. & Thelen, S. (2015). The Ontology-based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 29-39.

[34] Schmitz, P., Francesconi, E., Batouche, B., Landercy, S. P. & Touly, V. (2017). Ontological Models of Legal Contents and Users' Activities for EU e-Participation Services. *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 99-114.

[35] Bonatti, P., Kirrane, S., Polleres, A. & Wenning, R. (2017). Transparent Personal Data Processing: The Road Ahead. *International Conference on Computer Safety, Reliability, and Security*. Springer, 337-349.

[36] Wilson, R. (2004). *The Role of Ontologies in Teaching and Learning*. TechWatch Reports.

[37] Wakeford, R. (2002). Principles of Student Assessment. In Fry, H., Ketteridge, S. & Marshall, S. (Eds.). *A Handbook for Teaching and Learning in Higher Education. Enhancing Academic Practice*, 2nd ed., Routledge, 51-67.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190007

Automation, Legislative Production and Modernization of the Legislative Machine: The New Frontiers of Artificial Intelligence Applied to Law and e-Democracy

Gianluigi FIORIGLIO
Sapienza Unviersità di Roma - Dipartimento di Scienze Politiche (Italy)

Abstract. Electronic democracy is still far from being realized and several issues must be solved in order to make it possible. The quantitative problem of popular participation is one of them, but it can be mitigated through automation. This Chapter proposes two main applications that may help building a multilevel digital *agora* where *demos*, lawmakers, governments, and public administration may cooperate. The first is related to the integration, in each platform used for this purpose, of specific decision support systems. The second is inherent in the use of IT tools that, integrated into a digital *agora*, allow to transform the multiplicity of individual contributions into a general will.

Keywords. e-democracy, lawmaking, artificial intelligence, digital agora

1. Introduction

‘Electronic democracy’ (or e-democracy or digital democracy) may be studied from various viewpoints: law, sociology, computer science, philosophy, etc. However, there is no unambiguous notion of e-democracy; moreover, sometimes technology is seen as harmful for democracy and other times as a panacea for its problems. Thus, we can look at e-democracy ranging from exciting utopias to frightening dystopias, but both extremes are misleading. It is more useful to study e-democracy without forgetting that its core is ‘democracy’, not ‘electronic’. Thus, a previous in-depth analysis of its foundations as a democracy is needed to build upon a theoretical model of e-democracy. On the basis of such brief and preliminary observation, it may appear clear the perspective from which e-democracy is seen here, taking into account a model in which the technological component is secondary to the theoretical and conceptual ones.

Hence it could be defined as a new form of representative democracy in which tools of direct, deliberative, and participatory democracy are institutionalized; they can allow the exercise of popular sovereignty with advisory and legislative powers as appropriated.

Sovereignty is exercised in a digital *agora* (i.e. an electronic platform composed of a networked and multi-level system of smaller digital *agora*). This platform must be equipped with tools that simplify both the interaction of the *demos* and the various processes that are carried out, with the use of automatic data processing tools and support to the various stakeholders involved. Moreover, such platform may be designed in compliance with principles, values, and rules of each State in which it is implemented [1]. Several benefits could be achieved; for example, reducing the gap between politicians and the *demos*, empowering the latter.

This notion may be considered as a weak notion of e-democracy, because technology must adapt to each legal system, and not vice versa.

It is not possible to go deeper into the above-mentioned profiles, but it is clear that even such weak notion of e-democracy is still far from being realized notwithstanding the diffusion of new technologies and 'the rise of the Network Society', [2] in which digitalization is pervasive.

This is true even considering that technology plays a very important role in today's politics. On the one hand, the political debate is often carried out using electronic means: politicians and people discuss on social networks; political information is widely available; etc. On the other hand (and generally speaking), political and legislative procedures are still linked to the past. It means that some potential and important benefits of technology are not taken advantage of.

However, several issues must be solved in order to make e-democracy possible[1]. Even if it is not possible to investigate them in this Chapter, it is obvious that the quantitative problem of popular participation is one of them and it can be argued that it can be mitigated (although not solved) through automation. Automated acquisition and processing of information is an essential component for e-democracy: not only it can help overcoming the just mentioned issue, but it can also assist lawmakers in carrying out their task.

In fact, it is well known that, since the 1970s, the creation, diffusion, and technological development of computers and databases have allowed increasingly precise and refined processing of personal data exceeding the limitations inherent to those carried out with non-automatic tools. Similarly, the development and the use of advanced systems for processing particular types of information may enable the *demos* to be informed and effectively participate in the political life of the State, on the one hand, and lawmakers to fulfill their institutional tasks more efficiently, on the other hand. In other words, the legislative machine, regardless of its composition, can benefit from the automation of certain activities through the development and use of particularly advanced and 'smart' IT tools.

This Chapter proposes two main applications that may help building a multilevel digital *agora* where *demos*, lawmakers, governments, and public administration may cooperate.

The first is related to the integration, in each platform used for this purpose, of specific decision support systems, made up of a database of legal texts (laws and judgements) and of an expert system that allows people to understand the information stored in the database.

The second is inherent in the use of IT tools that, integrated into a digital *agora*, allow to transform the multiplicity of individual contributions into a general will.

[1] See *infra*, Section 2.

2. Some Issues of e-Democracy

It seems useful to briefly outline some issues of e-democracy before analyzing the problems and the possible solutions above mentioned because all these issues are connected; they are related to theory, design, and the *demos*.

The first issue concerns the weak theoretical framework of digital democracy, because in the past years the debate was focused on applications without taking into sufficient account the problem of its foundation[2].

As Janus Twofaced, the second major issue has two opposite profiles: as IT tools are not designed for e-democracy (but mainly for private uses), democracy is not designed for IT tools (and its fundamental legal texts have usually been written in a distant past, from a technological perspective). It is a matter of fact that political communities and discussions over the Web (and the social network) are often characterized by verbal fights (more than political debates), thus such processes may be more disruptive than aggregative; therefore, they are able to undermine the formation and realization of a general will, as in a dystopian model of electronic democracy. Digital democracy should instead be based on a custom-built platform technology whose functional and technical specification are set by each State in compliance with its rules.

The third issue regards the role of the *demos* in current democracies, that is especially delicate in a period in which the rise of populism could even suggest limiting such role. Broadly speaking, should be recommended that the *demos* participate in the legislative process? Yes: on the one hand, it could finally exercise its sovereignty in a continuous and not intermittent way; on the other hand, it could be held responsible because its choices would be much more effective and tangible than expressing dissent on social networks and electing its representatives (who can be blamed for their actions). The *demos* could become a real stakeholder in the political field and be responsible for its choices: it is easier to blame its representatives than itself.

However, the *demos* should be able to act and thus it is necessary to build a multilevel digital *agora*[3]. It may be designed as a permanent and highly organized computer assembly in which IT tools allow a continuous (and sometimes automated) interaction. Nonetheless, individual interactions should be processed to reduce complexity and recognize the general will, thus solving the quantitative problem of popular participation. This could also help in achieving a general and reciprocal listening, because the connection between speaking and listening is particularly weak on the Internet [6]; [7] and this may endanger e-democracy. Moreover, these tools should be controlled with no bias by public powers and with full transparency to avoid giving control of public infrastructures to private entities who could become the gatekeepers also of political procedures: today Google, Facebook, and few other entities already are the gatekeepers of political (and many other) information. For example, a search engine can decide what can be accessed

[2] Relatively few studies in the fields of legal philosophy, legal theory, political philosophy and political theory took into account digital democracy; however, this scenario is changing, and several contributions give life to an interesting debate. Among others, see: [1]; [3]; [4]; [5].

[3] The digital *agora* should not be unique: a network of *agora* should be provided to achieve effective multilevel governance and fully implement the principle of subsidiarity. Moreover, such reticular system is well suited to the traditional representative organs with a view to interparliamentary cooperation. The IT platform could be used to create *agora* between different States (for example, in the European Union). Each *agora* can adopt its own rules, but they could communicate using standard protocols and formats (as in the case of the World Wide Web, in which the most different devices are able to communicate regularly).

on the Web; it can hide and/or rank information making it hardly be found: "By controlling the communication infrastructure of the Internet, they have become information gatekeepers"[4] [8].

3. Artificial Intelligence, Law, and e-Democracy

Before analyzing the proposed decision support systems and the IT tools tailored for the digital *agora*, it is necessary to dwell on issues related, in general, to artificial intelligence applied to law and, more specifically, on how it could be used in the field of e-democracy; then risks and benefits will be analyzed in the perspective of the aforementioned disciplines.

First of all, the application of artificial intelligence to law is one of the most traditional and stimulating fields of investigation of legal informatics, but also one of the more difficult.

So far, the dream, or nightmare, of an electronic judge was not realized as it was feared in the Sixties[5], but there is no doubt that considerable progress has been made towards technologies that allow execution of tasks that require 'intelligence' and that, in a more or less distant future, will make possible this and other objectives that today seem unattainable (raising the question about the opportunity of using similar means).

However, the law still is an extremely difficult area for information technology. In the recent past, the problem of semantic understanding has been connected to the fact that a computer cannot understand what has not been algorithmically formulated. Obviously, not all the laws can be translated into algorithms; moreover, the formalization of the law – and therefore a strictly literal interpretation – cannot be taken to the extreme. In fact, to achieve a rational thought, rules must be followed; in principle, they are incompatible with those unhappy, contradictory and incomprehensible expressions contained in many legal texts. This, on closer inspection, applies to the language *tout court*, which is often vague, indeterminate, and equivocal.

Despite such difficulties, IT studies are making considerable progress and increasingly sophisticated algorithms are being developed; natural language is gradually better understood by automated machines and tools. After all, even voice assistants are a 'common technology'[6].

In particular, the ability to manage increasingly complex and numerous information, as well as to direct one's conduct on their basis (and on the modifications of the digital and material environment), is a realization of cybernetics (and therefore of the interdisciplinary theories of Norbert Wiener [10]), on the one hand, and of the three areas of investigation of jurimetrics outlined in the Sixties, on the other hand. More specifically,

[4] Moreover, "Users have become dependent on search engines, viewing them as authoritative and reliable. Search engines have become the tools through which the democratic potential of the Internet can be advanced or hindered" [8] p. 145.

[5] However, see [9]: the Authors have built a model designed to predict the behavior of the Supreme Court of the United States in a generalized, out-of-sample context. They developed a time evolving random forest classifier which leveraged some unique feature engineering to predict more than 240,000 justice votes and 28,000 cases outcomes over nearly two centuries (1816-2015). The authors used only data available prior to decision and, over nearly two centuries, they achieved 70.2% accuracy at the case outcome level and 71.9% at the justice vote level.

[6] Just think at Amazon Alexa, Apple Siri, and Google Assistant.

these areas are related to the archiving and electronic retrieval of legal information, the forecasting of future judgments of the judicial authority through the analysis of rulings and, finally, to the application of logical criteria to legal issues to logically control the reasoning that led to the judicial decision.

In the following decades the desired results were not achieved; technologies were not advanced enough, and judges had to evaluate not only previous rulings but also the evolution of the society (and such evaluations may lead to different interpretations).

The same happened with the studies on artificial intelligence, which have received maybe too criticism without recalling some of their indisputable achievements in the understanding of intelligence, in the development of new formal tools and new computer architectures; yet, IT development has led, at the beginning of the 21st century, to a progressive diffusion of hardware and software tools based on more or less refined techniques and applications of artificial intelligence.

We can draw important lessons at a theoretical level from these early theoretical and practical experiences, especially if we make reference to studies on legal philosophy and theory.

First of all, we can obtain the necessary theoretical bases for a potential e-democracy platform by updating the three areas of investigation of jurimetrics [11].

Today the information field is much more evolved; this is crucial because it allows any person (despite being a citizen) to acquire knowledge of both the regulatory framework and the relevant factual data, while being able to use automated filters.

The logical-decisional field can be understood in the sense of a system of decision support, not being able to replace the human one, as will be argued below.

Finally, the forecast field is necessary for a possible *ex post* evaluation of a potential decision, even if this tool is particularly sensitive because it may be abused and therefore should be limited only to positive law (with only reference to the regulatory impact).

4. Building a Digital Agora: Decision Support Systems and (New) IT Tools

A digital *agora* requires tools to acquire and make available information to the *demos*, on the one hand, and to let it take decisions. The former will feed a database, while the latter will provide means to express *demos*' opinion and organize the multiplicity of different manifestations of people's will, thus overcoming the quantitative problem of popular participation.

Particularly, the digital *agora* (and all of its nodes) should include specific decision support systems, made up (at least) of a database of legal texts (laws and judgements) and of an expert system that allows people to understand the information stored in the database. The aforementioned systems should be usable by citizens (or legitimate users) in order to enable them to fully and consciously participate in a digital *agora* thanks to the possibility of forming their opinion on a particular subject or case with the intermediation of a subject whose automated impartiality is guaranteed and verifiable.

The databases should be arranged as a network, like the Internet and the Web. They must be interoperable, so that they can: (i) connect to each other; (ii) be accessed by each digital agora; (iii) be consulted by each citizen or by any legitimate user[7].

[7] The limits of digital citizenship may be less strict than the ones of traditional citizenship.

Although these databases should be public[8], the progress made in the provision of these services by private entities can be particularly useful to drive evolution of public ones in data collection, record, organization, and processing, on the one hand, and being queried and presenting results, on the other hand. After all, everybody knows that in the Information Society there is a multitude of databases: not by chance, the most precious resource is today the information itself. Even when they contain billions of data, they can be extremely efficient and be accessed in response to billions of users' queries using several devices (smartphones, tablets, computers, cars, etc.).

However, information retrieval systems may have bugs or be not efficient, for example partially or totally hiding certain information, or be wrong in ranking results of users' queries[9].

Notwithstanding some bugs (and the potentially critical impact for those who are damaged by them), these platforms are increasingly efficient and precise, but they are provided by private entities executing secret algorithms and computer codes, so that they can be studied only on the basis of their outputs.

Even if these obstacles can be overcome in providing a digital *agora*, a cooperation between public and private sectors could be very useful. In fact, the public sphere is characterized by many databases managed by public administrations that progressively, albeit slowly, become interoperable[10], but archiving and retrieval of information can be long and complex activities; moreover, there is the problem of understanding texts and retrieved data, which is accompanied by their uselessness for practical purposes unless the individual is not one of the stakeholders of a given process (legislative, regulatory, judicial, etc.) and therefore has a concrete interest. Without a guide to make the ever-increasing amount of information available online, how can we hope that, even in the hypothesis in which the *demos* is called to participate actively, is it really able to do so?

Since it is important to make the *demos* both able to access and to understand information in order to actively participate in the public sphere, information technology can strongly contribute by elaborating complex information and making it relatively simple.

Thus, databases of several types should be used and include not only the legal ones. In fact, the amount of information already memorized by each State could be complemented by those freely available on the Net and managed by private parties. The joint analysis of this huge amount of information can be an essential resource for electronic democracy, also through automated analysis using Big Data and machine learning technologies that must be carried out in compliance with the fundamental principles of each democracy.

In fact, while the process of dematerialization of legal sources – which is often accompanied by the provision of free access to it – is growing, it is not enough. It would be necessary to provide advanced databases in which specific expert systems process information about a particular sector and make it usable for the *demos* who has to exercise its sovereignty in the digital *agora*.

[8] Generally speaking, a recent report on how UK media cover AI reveals that a majority of articles are pegged to industry concerns, and a plurality of sources originate from industry. The role of public action in addressing AI is usually undermined, while coverage frequently amplifies self-interested assertions of AI's value and potential and positioning AI primarily as a private commercial concern [12].

[9] Web search engines are an emblematic example of what has just been argued; see [13].

[10] See, among others, [14]. In particular, the Authors make a distinction between technological data, human, and institutional interoperability (p. 5-7).

Therefore, the contribution of legal informatics is fundamental, since the legal databases can be profitably organized thanks to the studies carried out on this subject for several decades and this organization can be a guide for a better implementation of the automatic systems for the understanding of law. More generally, the aforementioned databases could certainly be useful also for traditional lawmakers.

In other words, the information system must be designed taking into account not only mere indexation aimed at retrieving data, both legal and not, but also selection by association. Under this regard, it may be useful to actualize a pioneering thesis of Luigi Lombardi Vallauri by correlating the democratic nature of legal information with the democratic nature of the legal system, since there can be no political obligation if the content of the obligation cannot be known, or it is too difficult to be known [15]. The contribution of legal informatics is needed to create a system of legal information oriented to the citizen who not only offers him or her the 'simple' legal documentation but also facilitates his or her understanding ('translating it') and also provides him or her with links to set up or even to solve a specific legal problem.

This thesis is very suggestive and its application to the field of e-democracy is very useful, especially if we argue that knowledge of law is not equivalent to knowledge of legal information, for which mediation cannot be overcome [16]. We can therefore look at the thesis of Lombardi Vallauri both taking into account this point and claiming the possibility of achieving the creation of a 'legal automaton' capable of providing an opinion: this would provide both the elements necessary for the democratic nature of legal information, i.e. documentation and opinion. If, however, this does not yet seem feasible (and maybe undesirable for someone), it can be said that such a system can be better adapted to the field of electronic democracy than to democracy *tout court*, because a network of databases in which several software agents act would be supported by human action for several sectors (law, economics, etc.). Thus, it would be possible to set up a truly usable system for each citizen or individual qualified to interact in the digital *agora*.

Therefore, alongside a system that makes specialist information available and that 'translates' it into understandable terms, as far as possible, it is necessary to put an expert system. These tasks can be performed by automated software which, on the one hand, builds and feeds each database, then performs the necessary processing, and, on the other hand, assists each individual who queries it.

These tools can be evolved to aid achieving the democratic ideal.

This system should be accompanied by a second set of IT tools that, integrated into a digital *agora*, should be aimed at allowing to transform the multiplicity of individual contributions into a general will; this may happen through a gradual reduction of the number of such interactions thanks to automated means. Moreover, using the same technologies used for the first application, it is possible to simulate, in advance, the effects of one or more decisions (just think of the participatory budget, in which each decision can influence the others).

In detail, in each digital *agora* it is possible to assume several phases of discussion that must end within a specified time. At the end, the will of the participants is 'interpreted' using sentiment analysis techniques to synthesize them and reduce the field of discussion before voting. In this way it is possible to go beyond the limits of the binary logic of voting and reach a compromise between the discussion by a multiplicity of subjects and the need to reach the decision-making phase: both are necessary in any democratic process. This perspective, therefore, is based on the paradigm of automation.

However, in the perspective of digital democracy, what logic should be followed by the digital *agora* platform and what relationship should it have with truth? An e-democracy platform cannot ascertain the truth of the premises used in each reasoning, even if they can allow a reasonable degree of truth in relation to specific types of data where automated tools are provided (if a cross check is made possible). In other words, data entered into the system or acquired by its own software agents will constitute the premises of each machine reasoning, but its correctness should be evaluated upstream of the computer system (for example, with regard to the correct insertion of legal texts or to the number of citizens officially residing in a municipality). Thus, some databases must be characterized by absolute transparency and be constantly and fully available online, as often happens in many countries concerning the legislative texts and the work of the parliaments. In others it will be necessary to continue to reconcile the need for transparency with the right to privacy of any subjects involved or at least other interests actually involved.

Moreover, this electronic platform must be continuously nourished by new knowledge that is not exclusively juridical; however, today even a hyperintelligent automaton could not have that vast sphere of experience which only a human being possesses, and which is constantly growing (but, as the Information Society grows stronger, this gap reduces)[11].

5. Conclusions

It can be concluded that a decision-making system in the legal field that is completely autonomous could not be satisfactory, even if fed by knowledge deriving from studies and research as well as from machine learning. As e-democracy can be an important support to complement traditional democracy, so automation in the legal field can be a precious help and probably irreplaceable in the future for the various legal operators.

Furthermore, the whole legislative machine can benefit from the benefits of automation through the development of software that can assist the various parties involved in the whole process. The demos and the traditional lawmakers could benefit from tools that allow them to understand the factual situation, the current legislative framework, and the potential framework for the outcome of the new rules to be discussed or the administrative decisions to be taken. The legal operator can obtain an advanced support that helps her or him to orient himself in the multiplicity of regulations in force (at several levels), but, in any case, subject to his professional and particularly qualified judgment. More generally, both the governments and the *demos* could benefit from observing the reality also by making reference to Big Data, provided that certain criteria are respected: transparency, neutrality and objectivity (if and where possible, and taking pluralism into account).

[11] The applications discussed in this paragraph should be designed by experts of legal informatics, in cooperation with computer scientists, and with the fundamental contribution of experts of legal philosophy, legal theory, and constitutional law. This is due to the fact that its framework must be built taking into account a specific legal system.

However, both philosophical and positive law problems arise, also with reference to e-democracy. In fact, automation is the key to overcoming quantitative issues that make impossible to realize today a democracy like the Athenian democracy one, and which entail the excessive burden of certain proposals aimed at involving the population (such as deliberative polls). But it can also be used to control and manipulate the *demos*. It is therefore necessary to look towards a *constitutionally oriented* automation: it should be always respectful of the fundamental principles generally recognized in constitutional democracies. Nevertheless, it is not an easy task due to many factors, including the secrecy surrounding computer codes and algorithms, as well as the partial predominance of (private) technological powers over others, including public ones.

References

[1] Fioriglio, G. (2017). *Democrazia elettronica. Presupposti e strumenti*. Cedam-Wolters Kluwer.

[2] Castells, M. (2009). *The Rise of the Network Society. The Information Age: Economy, Society, and Culture*, 2nd ed., Wiley-Blackwell.

[3] Gometz, G. (2017). *Democrazia elettronica. Teoria e tecniche*. ETS.

[4] Mindus, P. (2014). What Does E- Add to Democracy? Designing an Agenda for Democracy Theory in the Information Age. In Bishop, J. (Ed.), *Transforming Politics and Policy in the Digital Age*, IGI, 200-223.

[5] Prins, C., Cuijpers, C., Lindseth, P. L. & Rosina, M. (Eds.) (2017). *Digital Democracy in a Globalized World*. Edward Elgar Publishing.

[6] Hindman, M. (2009). *The Myth of Digital Democracy*. Princeton University Press, 16-17.

[7] Benkler, Y. (2006). *The Wealth of Networks. How Social Production Transforms Markets and Freedom*. Yale University Press.

[8] Laidlaw, E. B. (2008). Private Power, Public Interest: An Examination of Search Engine Accountability. *International Journal of Law and Information Technology, 1*, 114.

[9] Katz, D. M., Bommarito, M. J. & Blackman, J. (2017). *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, https://ssrn.com/abstract=2463244.

[10] Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*, MIT Press.

[11] Loevinger, L. (1949). Jurimetrics. The Next Step Forward. *Minnesota Law Review, 33*(5), 455-493.

[12] Brennen, J. S., Howard, P. N. & Nielsen, R. K. (2018). *Industry-Led Debate: How UK Media Cover Artificial Intelligence*, Reuters Institute for the Study of Journalism, University of Oxford, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-12/Brennen_UK_Media_Coverage_of_AI_FINAL.pdf.

[13] Fioriglio, G. (2015). Freedom, Authority and Knowledge On Line: The Dictatorship of the Algorithm. *Revista Internacional de Pensamiento Político, 10*, 395-410.

[14] Palfrey, J. & Gasser, U. (2012). *Interop. The Promise and Perils of Highly Interconnected Systems*. Basic Books.

[15] Lombardi Vallauri, L. (1975). Democraticità dell'informazione giuridica e informatica. *Informatica e diritto*, 1, 3-10.

[16] Tincani, P. (2015). La stanza cinese, i database, l'esperto. In Perri, P. & Zorzetto S. (Eds.), *Diritto e linguaggio. Il prestito semantico tra le lingue naturali e i diritti vigenti in una prospettiva filosofico e informatico-giuridica*. ETS, 145-148.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190008

Reasoning with Deontic Notions in a Decidable Framework

Enrico FRANCESCONI
Publications Office of the European Union; Istituto di Teoria e Tecniche dell'Informazione Giuridica – ITTIG-CNR (Italy)

Abstract. The development of legal reasoning using decidable fragments of knowledge modeling languages is essential in the Semantic Web for the huge amount of triples available nowadays as Linked Open Data. This Chapter introduces a framework for legal knowledge representation and reasoning based on the distinction between the concepts of provision and norm, suited for different kinds of legal reasoning: legal provisions accessibility and norm compliance, respectively. The proposed framework allows the addressed types of reasoning to be implemented using OWL 2 decidable profiles and reasoners. Examples of decidable reasoning within the proposed framework are presented and tested.

Keywords. legal reasoning, legal provisions retrieval, norm compliance, semantic web, description logic

1. Introduction

The transformation of legal regulations in machine readable, actionable rules represents a precondition for developing systems endowed with automatic reasoning facilities for advanced information services in the legal domain.

In literature several approaches have been proposed aiming to formalize legal rules able to support automatic reasoning, as reasoning on deontic notions [1], reasoning for norm compliance [2] or for legal argumentation [3]. When implemented in the Semantic Web, legal reasoning approaches usually utilize languages like OWL/RDF(S) for modeling real world scenarios, as well as mainly SWRL, RIF or LegalRuleML for representing legal rules for such scenarios. Approaches for modeling real world scenarios and rules often, even not always [4], result in non-decidable profiles, so that the available reasoners are not guaranteed to be tractable from a computational point of view in large scale.

The current successful trend of Semantic Web implementation according to a Linked Open Data approach has produced, and is supposed to produce in the next few years, a huge and growing amount of RDF triples, representing concepts, legal rules and facts. The availability of decidable reasoners is therefore essential for dealing with the huge amount of Linked Open Data (LOD), so to guarantee the computational tractability of reasoning. Hence the need to represent the semantics of LOD triples by decidable fragments of knowledge modeling languages.

Also for these reasons in [5] a study has been carried out about how the Linked Open Data framework (including RDF(S)/OWL, LegalRuleML and SPARQL) can be applied to the formalization, publication and processing of legal knowledge, in particular normative requirements and rules.

In terms of knowledge modeling languages for the Semantic Web, OWL DL represents a decidable profile (subset) of OWL, and OWL 2 extends the OWL expressiveness introducing three additional decidable sub-profiles:

- OWL 2 DL: direct extension of OWL DL within the OWL 2 semantics, thus representing a decidable profile of the First Order Logic for OWL 2;
- OWL 2 EL: particularly useful in applications employing ontologies that contain very large numbers of properties and/or classes. It introduces specific restrictions allowing for example existential quantifications, while universal quantification or cardinality restrictions are not allowed;
- OWL 2 QL: particularly useful in applications dealing with a large amount of data. In this profile for example existential quantification to a class expression or a data range are not allowed;
- OWL 2 RL: particularly useful in scalable reasoning without sacrificing too much expressive power. It supports all axioms of OWL 2 apart from disjoint unions of classes (DisjointUnion) and reflexive object property axioms (ReflexiveObjectProperty).

In this Chapter we introduce a legal reasoning framework based on the distinction between the concepts of *Provision* and *Norm*, suited for different kinds of legal reasoning: legal provisions accessibility and norm compliance, respectively. Moreover, an approach based on decidable OWL 2 profiles is presented and tested. The claim of this study is not to address the whole complexity of legal reasoning, including for example non-monotonic reasoning, resolution of norm conflicts [6] or reasoning with incomplete and contradictory information, for which reasoners exist [7]; [8]. This study rather aims to present an approach which can be effectively implemented in a decidable framework for the type of reasoning mentioned (provisions retrieval and norms compliance). On the other hand this study may create the ground for investigating how more complex legal reasoning types actually affects the computational burden of the present approach.

This Chapter is organized as follows: in Section 2 a review of related work about legal reasoning is given, including examples within a description logic computational complexity; in Section 3 the distinction between the concepts of *Provision* and *Norm* from the legal theory point of view is discussed [9]; in Section 4 an approach for modeling deontic notions and implementing Hohfeldian reasoning for legal provisions retrieval within a decidable computational framework is recalled [10]; [11] and tested on examples; in Section 5 an approach for modeling norms using ontologies, able to implement norm compliance checking within a decidable computational framework is described and tested on examples; in Section 6 some conclusions are reported.

2. Related Works

OWL is the state-of-the-art standard for knowledge modeling in the Semantic Web, effectively used for creating ontologies able to represent concepts and relations of real world scenarios [12]: examples in the legal informatics literature are LRI-Core [13], LKIF [14], CLO [15], DALOS [16], PrOnto [17]. On the other hand legal rules in the Semantic Web have been represented in literature using a variety of languages. [18] proposed to use SWRL or RIF in combination with an ontological representation of norms and facts; [19] introduced rules description using specific XML schemas in combination with ontologies. More recently LegalRuleML [20] as specialization of the RuleML standard has been proposed as language for representing legal rules.

In the last few years several studies have been made to approach the problem of legal reasoning in a decidable computational complexity profile. [21] proposed HARNESS, a knowledge-based system developed within the ESTRELLA project, able to implement reasoning on norms for legal assessment, basically the evaluation whether a case is allowed or disallowed given an appropriate body of legal norms. HARNESS includes two knowledge bases: a domain ontology and a set of norms representing the normative articles. Both are developed within the OWL 2 DL profile. Assessment consists of classifying the individuals and properties making up a case description in terms of both ontology and norms simultaneously. The result tells whether norms are violated or not.

As anticipated in Section 1, recently in [5] an approach to represent legal rules as Linked Open Data has been proposed. It aims to respond to the requirements of representing and reasoning on the deontic aspects of normative rules with standard Semantic Web languages, as RDFS and OWL for knowledge representation and SPARQL for inquiries. The rationale of the proposed approach is the coupling of OWL reasoning with SPARQL rules to formalize and implement reasoning (as for example deontic reasoning). In particular normative requirements are represented using LegalRuleML and the the deontic conclusions of the legal rules are added to each named graph of the concerned state of affairs.

In [22] an OWL 2 judicial ontology library (JudO) representing the interpretations performed by a judge when conducting legal reasoning towards the adjudication of a case is illustrated. On the other hand [23] combines the features of description logic-based ontologies with non-monotonic logics such as defeasible logics.

In this Chapter we present an approach based on the distinction between *Provisions* and *Norms* represented by decidable fragments of OWL 2/RDFS for knowledge modeling and rules representation. This allows us to rely on available decidable reasoners able to derive implicit knowledge and conclusions on the model and related individuals, while leaving to SPARQL the sole task to query the dataset of inferred triples in order to verify deontic conclusions or norm compliance.

3. Provisions and Norms

According to the legal theory point of view, the legal order can be seen as a legal discourse composed by linguistic entities or *speech acts* [24] with descriptive or prescriptive functions. Every linguistic entity can be seen in a twofold perspective: as a set of signs organized in words and sentences, as well as the meaning of such signs. Following the same twofold view for the legal domain, we can distinguish two levels of interpretation of a linguistic entity expressing a legal rule: in terms of a set of signs organized in words and sentences for creating a normative statement, typically called *Provision* [25]; [26], and in terms of the meaning of such normative statement, typically called *Norm* [27]; [9].

Provisions have been classified in [26] in terms of provision types, organised into two main groups (Figure 1): *Rules* and *Rules on Rules*. *Rules* can be *Constitutive Rules*

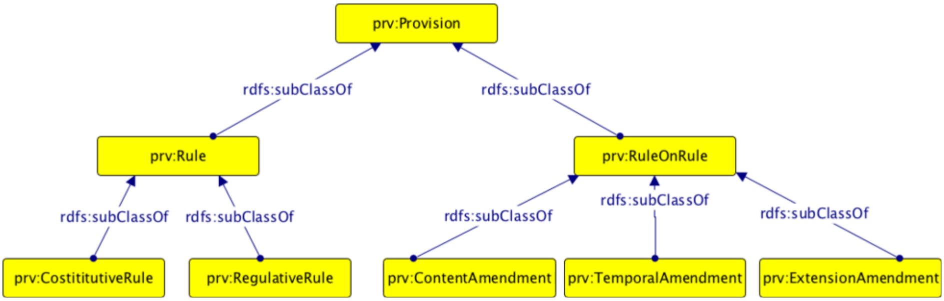


Figure 1. The *Provision Model* top classes ('prv:' is the namespace for provisions)

as *Definition* introducing entities, or *Regulative Rules* as the deontic concepts *Duty*, *Right*, *Power*, etc., regulating subject roles and activities. *Rules on Rules* are different kinds of amendments: *Temporal*, *Extension* or *Content amendments* as *Insertion*, *Substitution*, *Repeal*. Each provision type is characterized by specific *attributes* (for example the *Bearer* or the *Counterpart* of a *Right*), reflecting the lawmaker directions. Provision types and attributes can be considered as a sort of metadata model able to analytically describe fragments of legislative texts, hence the name of *Provision Model* [28]. In Figure 2 the *Provision Model* classes prv:Right and prv:Duty, and the related properties (as prv:hasRightBearer, prv:hasDutyCounterpart, etc.), under the namespace prv: for provisions in the Semantic Web, are sketched.

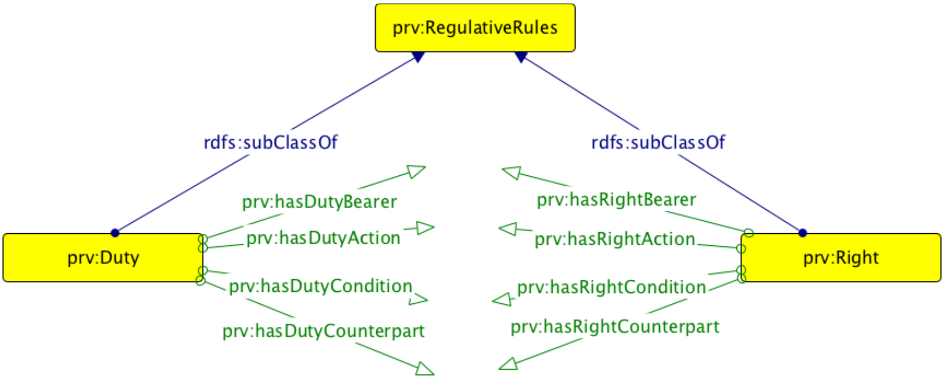


Figure 2. Examples of provisions attributes, represented as ontology classes and properties

Norms represent the application of provisions; as such they represent the product of an interpretative process [29].

Provisions and related norms have different roles and properties pertaining to the different abstraction levels they operate at. The need of distinguishing between provisions and norms becomes essential when we observe that there may be not a bijective relationship between them: a norm can be expressed by different provisions, as well as it can be valid the opposite, namely one provision can include more norms [30].

Moreover, they have different relationships with time. Provisions, as pure textual objects, are the product of lawmaking (legal drafting activity and promulgation) and they have a specific relation with time represented by the *in-force* date, namely the starting date of its existence in the legal order. On the other hand norms are the meaning of provisions, namely their interpretation; as such they have a specific relation with time represented by the *efficacy* date, namely the starting date in which a norm can be concretely applied. Therefore, while it is obvious that we can have cases of provisions not in-force and related norms not effective, as well as provisions in-force and related norms effective, we can also have provisions in-force and related norms not effective, as well as the symmetric case (this last one is usually referred as *ultra-activity* of a norm).

Having different nature, such concepts operate in different domains. A provision, as pure textual object, represents the building block of the legal order (new provisions can enter or leave the legal order). On the other hand norms, can either modify the text of other provisions (in case of different type of amendments) or can introduce restrictions on the real world (in case of obligations, for example).

We have underlined the different nature of such objects as a background for introducing an approach for legal reasoning using such different concepts in different types of legal reasoning. Advanced legal document retrieval systems able, for example, to implement Hohfeldian reasoning on deontic notions, is a type of reasoning managing textual information, thus pertaining to provisions. Legal compliance checking is a process aiming to verify if a fact, occurring in the real world, complies with a legal norm. Real world scenarios and facts can be effectively represented in terms of ontologies and related individuals, respectively. Norms, which facts have to be compliant with, provide constraints on the reality, therefore they can be modeled, in particular as far as obligations are concerned, as restrictions on ontology properties, used for legal compliance checking.

Let's consider two examples of rules, R1 and R2, to illustrate our approach:

R1 : *The supplier shall communicate to the consumer all the contractual terms and conditions*

R2 : *According to a [country] law one cannot drive over 90 km/h*

Both rules are speech acts, namely *Provisions* in specific regulations. In this sense R1 can be classified as a *Duty* of a *Supplier* towards *Consumer*, while R2 can be classified as an *Obligation* for any *Vehicle* in the related specific country. Such classification can be used for implementing an advanced retrieval system able to select provisions by provision types and attributes. Moreover, R1 is an example in which Hohfeldian reasoning can be applied, since that provision can be seen as a *Right* of the *Consumer* to receive communications by *Supplier*. Such provision has to be retrieved either if we search for the supplier's duties or if we search for the consumer's rights. In all these cases an advanced legal retrieval system endowed with legal reasoning facilities pertains to provisions.

When we consider the application of R1 and R2 on specific facts, we actually talk about *Norms*. As previously discussed, norms which facts have to be compliant with, can be viewed as constraints on the real world to be regulated.

In the case of R1, the scenario can be modeled in terms of an ontology including a class *Supplier*, having a boolean property *hasCommunicatedConditions*, while the norm R1 can be modeled in terms of restriction on that property, so that the compliant supplier must have set to 'true' the value of such property when he has complied with his communication duties.

Similarly, in case of R2, the vehicles circulation scenario of a specific country can be modeled in terms of an ontology including a class *Vehicle*, having a property *hasSpeed*, while the norm R2 can be modeled in terms of restriction on that property, so that the compliant individual vehicles must have set the value of such property in the interval [0.0, 90.0 Km/h] (detected for example by the police through a speed checker station). In both cases legal reasoning in terms of legal compliance checking can be obtained by managing norms modeled as restrictions on property values of ontology classes.

In Section 4 we recall the modeling approach, described in [10]; [11], about how legal reasoning on deontic notions (as reasoning over Hohfeldian relations) for an advanced legal provisions retrieval system can be modeled within a description logic implemented in OWL 2 DL. Similarly, in Section 5 we show how legal reasoning for norm compliance can be implemented by modeling norms as ontology properties restriction using decidable fragments of OWL 2. In both cases inferences and related legal reasoning can be effectively calculated using decidable reasoners.

4. Modeling Provisions for Advanced Legal Provisions Accessibility

In [11] and [10] it is shown how Hohfeldian relations on deontic and potestative notions can be managed within a description logic computational framework. We recall here the main aspects of the approach to show, for the examples R1 and R2, how *Provisions* can be used to implement an advanced legal provisions retrieval system, endowed with legal reasoning facilities, using a decidable fragment of OWL 2 (in particular OWL 2 DL), therefore exploiting existing decidable reasoners.

In this recall, we show the approach for deontic notions and their relations, sketched in the schema of Figure 3[1].

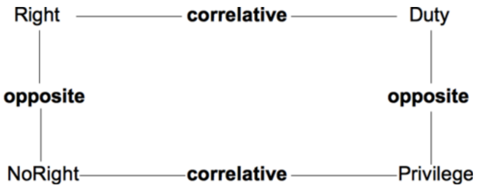


Figure 3. Hohfeldian relations on deontic concepts

In order to implement an advanced legal provisions retrieval system, it is necessary to describe the relations between provisions at the level of the Provision Model. For example the Hohfeldian relation between *Duty* and *Right* can be effectively represented by observing that a *Right*, in correlative correspondence with a *Duty*, is actually not explicitly expressed in the text, but represents an implicit provision, basically a different view of the *Duty* itself, where the values of the related bearer and counterpart attributes are swapped. Therefore, the Provision Model can be extended in terms of Duty and Right[2] implicit and explicit disjoint subclasses, able to represent a complete covering of the related superclass (ex: ExplicitRight and ImplicitRight disjoint subclasses represent a complete covering of the Right superclass).

Attributes can also be specified as regards both implicit and explicit provisions, so that hasImplicitDutyBearer and hasExplicitDutyBearer are sub-properties of hasDuty-

[1] More details on this modeling approach and its application to potestative notions (*Power, Liability, Disability Immunity*), can be found in [11] and [10].

[2] Where ‘prv:’, namespace for provisions, is hereinafter implied.

Bearer, as well as hasImplicitRightBearer and hasExplicitRightBearer are sub-properties of hasRightBearer.

To represent the hohfeldian fundamental relations between Duty and Right, firstly an equivalence relation between their explicit and implicit views is established: ImplicitRight ≡ ExplicitDuty and ImplicitDuty ≡ ExplicitRight. In Figure 4 the established sub-class and equivalence relations between Duty and Right in their explicit and implicit views are summed up.

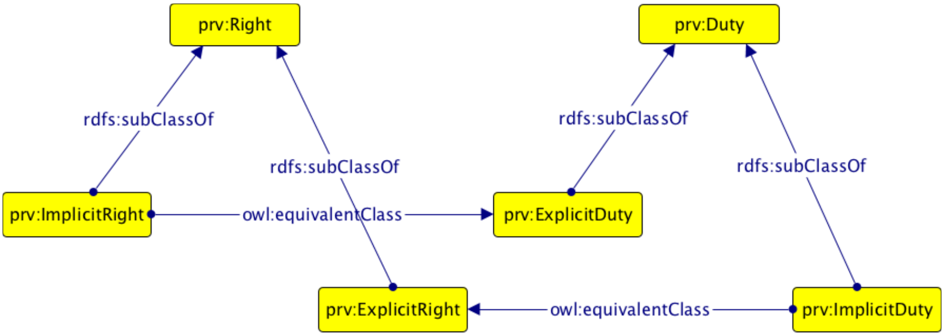


Figure 4. Sub-class and asserted equivalence relations between Duty/Right deontic correlative provisions

Moreover, equivalence relations between implicit/explicit Duty and Right attributes can be established. In Figure 5 the asserted properties of ExplicitDuty and ImplicitRight and their mutual equivalence relations are shown (hasImplicitRightBearer ≡ hasExplicitDutyCounterpart and hasImplicitRightCounterpart ≡ hasExplicitDutyBearer).

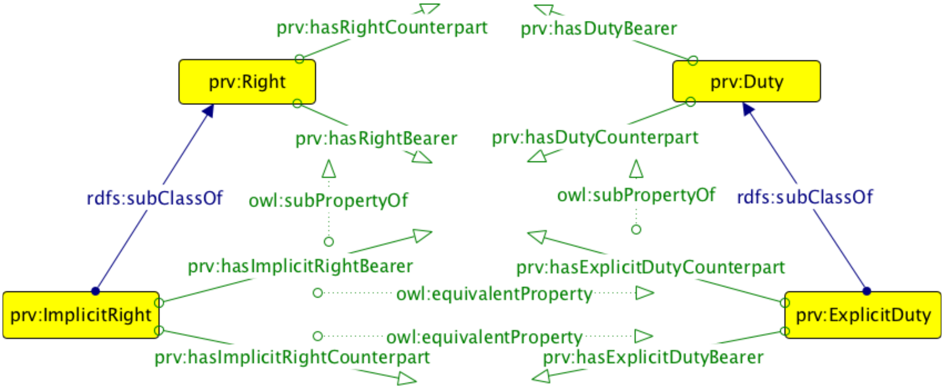


Figure 5. Asserted properties of ExplicitDuty and ImplicitRight and their mutual equivalence relations

The same holds for the asserted properties of ImplicitDuty and ExplicitRight and their mutual equivalence relations (hasImplicitDutyBearer ≡ hasExplicitRightCounterpart and hasImplicitDutyCounterpart ≡ hasExplicitRightBearer) (Figure 6) .

Note that the proposed patterns do not interfere with the relations between Right and Duty, which still hold. In fact, for the couple Right/Duty, an individual of ExplicitDuty is also an individual of Duty, given the axiom rdfs:subClassOf(ExplicitDuty,

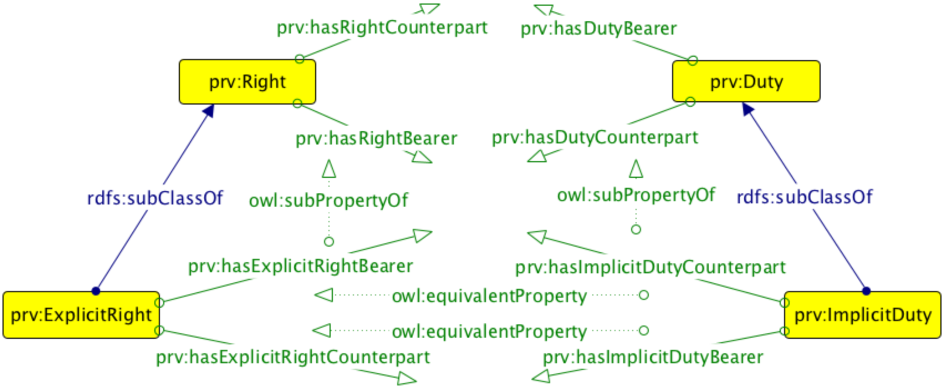


Figure 6. Asserted properties of ImplicitDuty and ExplicitRight and their mutual equivalence relations

Duty). Moreover the axiom owl:equivalentClass(ImplicitRight, ExplicitDuty) tells us that such individual is also an ImplicitRight, which is also a Right, given the axiom rdfs:subClassOf(ImplicitRight, Right). Since this is done symmetrically for explicit and implicit duties and rights, we can deduce that Right is equivalent to Duty, namely is another reading of the Duty itself, given that the union of the disjoint explicit and implicit subclasses covers completely the related superclass.

4.1. Provisions R1 and R2 representation and reasoning

Such modeling approach can be used to represent the provisions R1 and R2 in a way that a provision retrieval system can be implemented using a decidable reasoner. For example R1 can represented as follows:

```
<rdf:Description rdf:about="[URI]#R1">
 <rdf:type rdf:resource="prv:ExplicitDuty"/>
 <prv:hasExplicitDutyBearer rdf:resource="myo:Supplier"/>
 <prv:hasExplicitDutyAction
         rdf:datatype="http://www.w3.org/2000/01/rdf-schema#Literal">
         has communicated</prv:hasExplicitDutyAction>
 <prv:hasExplicitDutyObject rdf:resource="myo:Conditions"/>
 <prv:hasExplicitDutyCounterpart
         rdf:resource="myo:Consumer"/>
</rdf:Description>
```

where myo: is a fictitious namespace for a fictitious ‘MyOntology’, while the values of the provisions attributes can be either ontology concepts (as in the example for duty bearer and counterpart) or literals (as for the duty action). Note that only explicit provision classes (and consequently explicit properties) are used to annotate textual provisions, as they are the only provisions actually (explicitly) expressed in the text, while implicit provisions act as a sort of ‘abstract’ classes, which are used for reasoning.

As both the Provision Model (and related instances) result in OWL 2 DL profile, inferences can be calculated through an OWL 2 DL reasoner. In this example the Pellet[3] Java based OWL 2 DL reasoner is used to derive the inferred model.

[3] https://www.w3.org/2001/sw/wiki/Pellet.

Let's assume to query the system as follows in order to retrieve the suppliers' duties:

```
SELECT ?x WHERE  {?x  prv:hasDutyBearer myo:Supplier}
```

where x is the variable that will contain the identifiers of the retrieved provisions instances. In case the *non-inferred* model is queried, no provisions are retrieved since only ExplicitDuty and related attributes are used for provision annotation. In case the *inferred* model is queried, the inferred provisions are retrieved, either annotated as ExplicitDuty of Supplier, if any, or implicitly deduced by provision relations. By exploiting the established rdfs:subClass relations between provisions type and attributes, the system will act as virtually expanding the query in terms of hasExplicitDutyBearer and hasImplicitDutyBearer, thus being able to retrieve R1, which is the provision having Supplier as value of the property hasExplicitDutyBearer.

Similar considerations can be given if we want to retrieve the consumers' rights, as follows

```
SELECT ?x WHERE  {?x  prv:hasRightBearer myo:Consumer}
```

In this case hohfeldian reasoning is produced. In fact by exploiting the established rdfs:subClass and owl:equivalentClass relations between provisions type and attributes, the system will act as virtually expanding the query in terms of hasExplicitRightBearer and hasImplicitRightBearer. For the last one the following relation holds hasImplicitRightBearer $\equiv$ hasExplicitDutyCounterpart): this allows the system to retrieve R1, which is the provision having Consumer as value of the property prv:hasExplicitDutyCounterpart. Since this is the result of axioms established in the Provision Model for implementing hohfeldian relations, the result is an hohfeldian reasoning over provisions (namely searching for consumers' rights and retrieving the related suppliers' duties).

Very similar considerations can be given for the annotation of R2 and a query able to retrieve it. The provision R2 can be represented as follows:

```
<rdf:Description rdf:about="[URI]#R2">
 <rdf:type rdf:resource="prv:Obligation"/>
 <prv:hasObligationBearer rdf:resource="myo:Vehicle"/>
 <prv:hasObligationAction
         rdf:datatype="http://www.w3.org/2000/01/rdf-schema#Literal">
            cannot overcome 90Km/h</prv:hasObligationAction >
</rdf:Description>
```

where again, myo: is a fictitious namespace for a fictitious ontology "MyOntology", while the values of the provisions attributes can be either ontology concepts (as for the obligation bearer) or literals (as for the obligation action). The following SPARQL query is able to retrieve the provision R2:

```
SELECT ?x WHERE {?x  prv:hasObligationBearer myo:Vehicle}
```

Both the previous cases represent retrieval examples including legal reasoning on deontic notions which can be managed within the OWL 2 DL decidable computational profile.

5. Modeling Norms for Legal Compliance Checking

As discussed in Section 3, norms can be viewed as the application, subject to interpretation, of legal provisions, providing constraints on a real world scenario to be regulated. In the Semantic Web a real world scenario is usually represented by a domain ontology. In this context a norm, providing constraints to such scenario, can be modeled in terms of constraints on the domain ontology: for example, in case of obligations, as ontology property restrictions. In the following of this Section we show how the norms expressed in R1 and R2, can be represented as restrictions on the addressed scenarios, and how such a representation can be used for norm compliance checking, within a computational complexity decidable profile. Note that in the examples, the relations between text, provisions and corresponding norms have not been reported for the sake of simplicity, but can obviously be expressed by relations between the related URIs.

5.1. Norm R1 Representation and Compliance Checking

In the case of R1, the scenario can be modeled in terms of an ontology including a class Supplier, having a boolean property hasCommunicatedConditions. In OWL 2 terms the scenario concerning R1 can be expressed as follows:

```
<owl:Class rdf:about="myo:Supplier">
 <rdfs:comment
   xml:lang="en">The class of the Suppliers</rdfs:comment>
 <rdfs:label xml:lang="en">Supplier</rdfs:label>
</owl:Class>

<owl:DatatypeProperty rdf:about="myo:hasCommunicatedConditions">
  <rdfs:domain rdf:resource="myo:Supplier"/>
  <rdfs:range rdf:resource="xsd:boolean"/>
  <rdfs:comment
       xml:lang="en">The property describing purchasing
       conditions communicated or not</rdfs:comment>
  <rdfs:label xml:lang="en">has communicated the conditions
  </rdfs:label>
</owl:DatatypeProperty>
```

Norm R1, expressing a duty for the suppliers states that suppliers must communicate purchasing conditions to the consumers: the individuals of the class Supplier complying with this norm are all those ones belonging to the subclass SupplierR1Compliant identified by a restriction on the boolean property hasCommunicatedConditions to have value ‘true’ (Figure 7).

In other terms the norm R1 is represented as restriction on the property hasCommunicatedConditions able to identify the class SupplierR1Compliant which is equivalent (see owl:equivalentClass relation here below) to the class of all individuals for which the value of the property under consideration is ‘true’, as follows:

```
<owl:Class rdf:about="myo:SupplierR1Compliant">
  <owl:equivalentClass>
  <owl:Restriction>
   <owl:onProperty
         rdf:resource="myo:hasCommunicatedConditions"/>
```

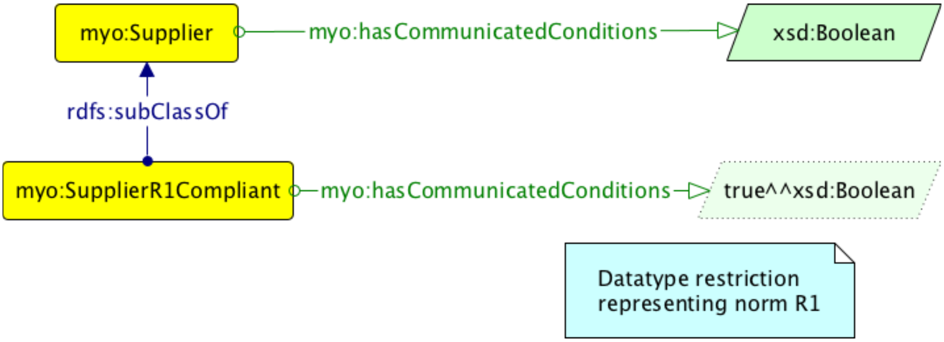


Figure 7. Norm R1 represented as restriction on the Supplier's property hasCommunicatedConditions (note that the subclass relation between SupplierR1Compliant and Supplier is inferred)

```
    <owl:hasValue
           rdf:datatype="xsd:boolean">true</owl:hasValue>
   </owl:Restriction>
   </owl:equivalentClass>
</owl:Class>
```

Such a representation for the real world scenario and related norm expressed by R1 results in the OWL 2 DL, as well as OWL 2 RL, decidable profiles.[4] This allows us to use a OWL 2 DL decidable reasoner, as for example Pellet, in order to implement reasoning facilities, preparing the ground for compliance checking with respect to R1. The inferred model produced by Pellet establishes a rdfs:subClassOf relationship between SupplierR1Compliant and Supplier (as shown in Figure 7), where SupplierR1Compliant is the class of all the individuals of type Supplier having 'true' as value of the property hasCommunicatedConditions. Therefore, compliance checking according to the norm R1 is a problem of checking if an individual of type Supplier belongs to the class SupplierR1Compliant.

As an example let's consider the following two individuals myo:s1 and myo:s2 of the class Supplier:

```
<myo:Supplier rdf:about="myo:s1">
 <myo:hasCommunicatedConditions rdf:datatype="xsd:boolean">
    false</myo:hasCommunicatedConditions>
</myo:Supplier>

<myo:Supplier rdf:about="myo:s2">
 <myo:hasCommunicatedConditions rdf:datatype="xsd:boolean">
    true</myo:hasCommunicatedConditions>
</myo:Supplier>
```

myo:s1 is an individual not compliant with R1, while myo:s2 is complaint with R1. The following SPARQL query

```
SELECT ?x WHERE { ?x rdf:type myo:SupplierR1Compliant }
```

[4] This can be verified using the Manchester validator at http://mowl-power.cs.man.ac.uk:8080/validator/.

is able to select the individuals which are complaint with R1 (in our case s2). Legal reasoning in terms of norm compliance checking is therefore performed within a decidable computational complexity profile.

5.2. Norm R2 Representation and Compliance Checking

In the case of R2, the vehicles circulation scenario can be modeled in terms of an ontology including a class Vehicle, having a datatype property hasSpeed with range in the xsd:float datatype. In OWL 2 terms, the vehicles circulation scenario concerning R2 can be expressed as follows:

```
<owl:Class rdf:about="myo:Vehicle">
  <rdfs:comment xml:lang="en">The class Vehicles
                                            </rdfs:comment>
  <rdfs:label xml:lang="en">Vehicle</rdfs:label>
</owl:Class>

<owl:DatatypeProperty rdf:about="myo:hasSpeed">
  <rdfs:domain rdf:resource="myo:Vehicle"/>
  <rdfs:range rdf:resource="xsd:float"/>
  <rdfs:comment xml:lang="en">Speed of a Vehicle
                                            </rdfs:comment>
  <rdfs:label xml:lang="en">has speed</rdfs:label>
</owl:DatatypeProperty>
```

Norm R2, expressing an obligation on the vehicles circulation scenario, states that, according to the related country law, one cannot drive over 90 km/h: the individuals of the class Vehicle complying with this norm are those ones belonging to the subclass VehicleR2Compliant having value ∈ [0.0, 90.0 Km/h] on the datatype property hasSpeed (Figure 8).

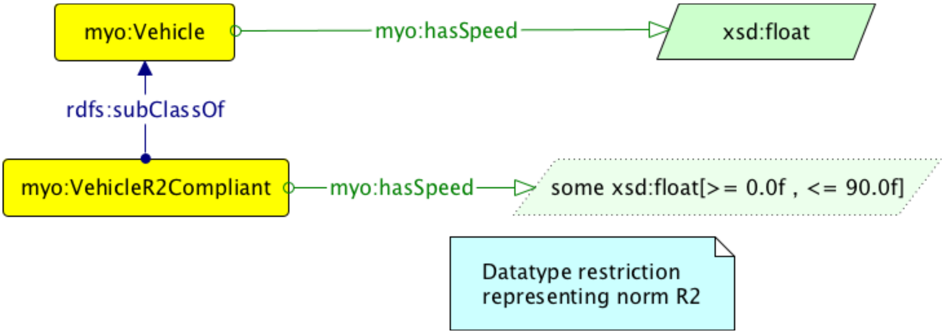


Figure 8. Norm R2 represented as restriction on the Vehicle's property hasSpeed (note that the subclass relation between VehicleR2Compliant and Vehicle is inferred)

In other terms the norm R2 is represented as restriction on the property hasSpeed able to identify the class VehicleR2Compliant which is equivalent to the class of the individuals for which the values of the property under consideration are in the range [0.0, 90.0 km/h]. In order to represent such constraints the following restriction on the datatype property myo:hasSpeed to values (inclusively) between 0.0 and 90.0 can be expressed by the xsd:minInclusive and xsd:maxInclusive datatype bound properties. In OWL 2 this results as follows:

```
<owl:Class rdf:about="myo:VehicleR2Compliant">
 <owl:equivalentClass>
 <owl:Restriction>
   <owl:onProperty rdf:resource="myo:hasSpeed" />
   <owl:someValuesFrom>
      <rdfs:Datatype>
      <owl:onDatatype rdf:resource="xsd:float"/>
      <owl:withRestrictions rdf:parseType="Collection">
      <rdf:Description>
        <xsd:minInclusive
           rdf:datatype="xsd:float">0.0</xsd:minInclusive>
      </rdf:Description>
      <rdf:Description>
        <xsd:maxInclusive
          rdf:datatype="xsd:float">90.0</xsd:maxInclusive>
        </rdf:Description>
      </owl:withRestrictions>
      </rdfs:Datatype>
   </owl:someValuesFrom>
 </owl:Restriction>
 </owl:equivalentClass>
</owl:Class>
```

Such a representation results in the OWL 2 DL decidable profile[5]. As in the previous example, the inferred model, produced by Pellet, establishes a rdfs:subClassOf relationship between VehicleR2Compliant and Vehicle (as shown in Figure 8), where VehicleR2Compliant is the class of all the individuals of type Vehicle having values of the property hasSpeed in the interval [0.0, 90.0 km/h]. Therefore, compliance checking according to the norm R2 is a problem of checking if an individual of type Vehicle belongs to the class VehicleR2Compliant.

As a concrete example, let's consider the following four individuals of the class Vehicle:

```
<myo:Vehicle rdf:about="myo:v1">
    <myo:hasSpeed
             rdf:datatype="xsd:float">50.0</myo:hasSpeed>
</myo:Vehicle>
<myo:Vehicle rdf:about="myo:v2">
    <myo:hasSpeed
             rdf:datatype="xsd:float">60.0</myo:hasSpeed>
</myo:Vehicle>
<myo:Vehicle rdf:about="myo:v3">
    <myo:hasSpeed
             rdf:datatype="xsd:float">70.0</myo:hasSpeed>
</myo:Vehicle>
<myo:Vehicle rdf:about="myo:v4">
    <myo:hasSpeed
             rdf:datatype="xsd:float">95.0</myo:hasSpeed>
</myo:Vehicle>
```

[5] This can be verified using the Manchester validator at http://mowl-power.cs.man.ac.uk:8080/validator/.

In this list, the individual myo:v4 is not compliant with R2 (having speed 95.0 Km/h $\geq$ 90.0 Km/h). The following query:

```
SELECT ?x WHERE { ?x rdf:type myo:VehicleR2Compliant }
```

is able to select the individuals which are complaint with R2 (in our case myo:v1, myo:v2, myo:v3). Also in this case the norm compliance checking is performed within a decidable computational complexity profile.

6. Conclusions and Future Developments

In this Chapter we have presented a legal reasoning approach based on the distinction between the concepts of provisions and norms, able to deal with different types of legal reasoning, in particular advanced legal provisions retrieval, as well as norms compliance checking. The method is based on the use of decidable fragments of OWL 2, able to guarantee the computational tractability of the approach. This represents an essential property of a legal reasoning system in the Semantic Web, characterized by a huge amount of Linked Open Data in the form of triples.

Actually, legal reasoning is characterized by a more complex logic, the typical case being the defeasible one. On the other hand, the aim of this approach is to identify a framework, which seems promising in terms of computational tractability, under whose umbrella investigating the sufficient conditions according to which some types of legal reasoning can be managed by decidable reasoners, so to guarantee not only a computational tractability, but also the possibility to reuse existing reasoners developed for decidable profiles of OWL 2.

In the next future we aim to explore the semantics and use of DL-safe Rules [31], which are a specific decidable fragment of SWRL, able to describe a type of rules where variables appears in the rule premise in both unary and binary predicates.

References

[1] Broersen, J., Condoravdi, C. & Shyam, N. (Eds.) (2018). *Deontic Logic and Normative Systems* - 14th International Conference, Deon 2018 (Utrecht, the Netherlands, 3-8 July 2018). College Publications.

[2] Muthuri, R., Boella, G., Hulstijn, J., Capecchi, S. & Humphreys, L. (2017). Compliance Patterns: Harnessing Value Modeling and Legal Interpretation to Manage Regulatory Conversations. In Keppens, J. & Governatori, G. (Eds.). *Proceedings of the 16th International Conference on Artificial Intelligence and Law, ICAIL*. ACM, 139-148.

[3] Prakken, H. & Sartor, G. (2015). Law and Logic: A Review from an Argumentation Perspective. *Artificial Intelligence, 227*, 214-245.

[4] Governatori, G., Hashmi, M., Lam, H., Villata, S. & Palmirani, M. (2016). Semantic Business Process Regulatory Compliance Checking Using LegalRuleML. In Blomqvist, E., Ciancarini, P., Poggi, F. & Vitali, F. (Eds.). *Knowledge Engineering and Knowledge Management*. Springer International, 746-761.

[5] Gandon, F., Governatori, G. & Villata, S. (2017). Normative Requirements As Linked Data. In Wyner, A. & Casini, G. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 30th Jurix Conference*. IOS Press, 1-10.

[6] Batsakis, S., Baryannis, G., Governatori, G., Tachmazidis, I. & Antoniou, G. (2018). Legal Representation and Reasoning in Practice: A Critical Comparison. In Palmirani, M. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 31st Jurix Conference*. IOS Press, 31-40.

[7] Lam, H. P. & Governatori, G. (2009). The Making of SPINdle. In Governatori, G., Hall, J. & Paschke, A. (Eds.). *Rule Interchange and Applications. RuleML 2009.* Lecture Notes in Computer Science. Springer, Vol. 5858.

[8] Antoniou, G., Dimaresis, N. & Governatori, G. (2008). A System for Modal and Deontic Defeasible Reasoning. In *Proceedings of the 2008 ACM Symposium on Applied Computing*. ACM, 2261-2265.

[9] Marmor, A. (2014). *The Language of Law*. Oxford University Press.

[10] Francesconi, E. (2014). A Description Logic Framework for Advanced Accessing and Reasoning Over Normative Provisions. *International Journal on Artificial Intelligence and Law, 22*(3), 291-311.

[11] Francesconi, E. (2016). Semantic Model for Legal Resources: Annotation and Reasoning Over Normative Provisions. *Semantic Web Journal. Special Issue on Semantic Web for the Legal Domain, 7*(3), 255-265.

[12] Casellas, N. (2008). *Modelling Legal Knowledge through Ontologies. OPJK: The Ontology of Professional Judicial Knowledge*. PhD Dissertation. Institute of Law and Technology, Autonomous University of Barcelona.

[13] Breuker, J. (2004). Constructing a Legal Core Ontology: LRI-core. In *Proceedings of the Workshop on Ontologies and Their Applications*, 115-126.

[14] Hoekstra, R., Breuker, J., Di Bello, M. & Boer, A. (2007). The LKIF Core Ontology of Basic Legal Concepts. *LOAIT, 321*, 43-63.

[15] Gangemi, A., Sagri, M.-T., Tiscornia, D. (2005). A Constructive Framework for Legal Ontologies. In *Law and the Semantic Web*. Springer, 97-124.

[16] Agnoloni, T., Bacci, L., Francesconi, E., Spinosa, P., Tiscornia, D., Montemagni, S. & Venturi, G. (2007). Building an Ontological Support for Multilingual Legislative Drafting. In Arno R. Lodder, A. R. & Mommers, L. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 20th Jurix Conference*. IOS Press, 9-18.

[17] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C. & Robaldo, L. (2018). PrOnto: Privacy Ontology for Legal Reasoning. In *Electronic Government and the Information Systems Perspective. EGOVIS 2018.* Lecture Notes in Computer Science. Springer, Vol. 11032, 139-152.

[18] Hoekstra, R., Breuker, J., di Bello, M. & Boer, A. (2009). Lkif Core: Principled Ontology Development for the Legal Domain. In Breuker, J., Casanovas, P., Klein, M. C. A., & Francesconi, E. (Eds.). *Law, Ontologies and the Semantic Web*. IOS Press, 21-52.

[19] Gordon, T. (2011). Combining Rules and Ontologies with Carneades. In *Proceedings of the 5th International RuleML2011@BRF Challenge*. CEUR-WS.org, Vol. 799.

[20] OASIS. (2017). *LegalRuleML Core Specification Version 1.0*, http://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/csprd02/legalruleml-core-spec-v1.0-csprd02.html.

[21] Van de Ven, S., Breuker, J., Hoekstra, R. & Wortel, L. (2008). Automated Legal Assessment in OWL 2. In Francesconi, E., Sartor, G. & Tiscornia, D. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 21st Jurix Conference*. IOS Press, 170-175.

[22] Ceci, M. & Gangemi, A. (2016). An OWL Ontology Library Representing Judicial Interpretations. *Semantic Web Journal. Special Issue on Semantic Web for the Legal Domain, 7*(3), 229-253.

[23] Ceci, M. (2013). Representing Judicial Argumentation in the Semantic Web. In *Proceedings of the 5th Workshop on Artificial Intelligence and the Complexity of Legal Systems (AICOL)*. Springer, 172-187.

[24] Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

[25] Raz, J. (1980). *The Concept of a Legal System*. Oxford University Press.

[26] Biagioli, C. (2009). *Modelli funzionali delle leggi. Verso testi legislativi autoesplicativi*. European Press Academic Publishing. Legal Information and Communications Technologies Series, Vol. 6.

[27] Guastini, R. (2010). *Le fonti del diritto. Fondamenti teorici*. Giuffrè.

[28] Biagioli, C. & Grossi, D. (2008). Formal Aspects of Legislative Meta-drafting. In Francesconi, E., Sartor, G. & Tiscornia, D. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 21st Jurix Conference*. IOS Press, 192-201.

[29] Kelsen, H. (1991). *General Theory of Norms*. Clarendon Press.

[30] Pino, G. (2016). *Teoria analitica del diritto. Chapter 2. Norma giuridica*. ETS, 144-183.

[31] Hitzler, P., Krötzsch, M. & Rudolph, S. (2009). *Foundations of the Semantic Web Technologies*. Chapman & Hall/CRC.

This page intentionally left blank

Part II

Challenges and Opportunities in Disseminating and Accessing Legal Information

Section II.1

Rules, Policies and Publication Models

This page intentionally left blank

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190010

The EU Council Conclusions on the Online Publication of Court Decisions

Marc VAN OPIJNEN [1]
Publications Office of the Netherlands (UBR|KOOP)

Abstract. The topic of public access to national court decisions has never been explicitly on the agenda of the Council of the European Union, since it has been viewed upon as a national responsibility. Recent developments though – like the introduction of the European Case Law Identifier (ECLI) and the go-live of the ECLI search engine, operated by the European Commission and containing millions of national court decisions – have raised awareness about the importance of online accessibility of national case law for the European legal order. This has set the stage for the adoption, on 8 March 2018, of the 'Conclusions of the Council and the Representatives of the Governments of the Member States Meeting within the Council on Best Practices regarding the Online Publication of Court Decisions'. This Chapter discusses various aspects of these Conclusions. First of all, the character of such Council conclusions as a soft law instrument will be explained. Secondly, the document is reviewed in a broader context of recent policy developments and other (semi-) legal instruments. Finally, the substantive contents of the document will be examined. Although most of the best practices prescribe what is already common practice in all or most EU Member States, some provisions call upon governments and judiciaries to implement strategies that are not commonplace yet, e.g. to supply for some kind of importance qualification, indicating which, and to which extent court decisions are of relevance for others than the parties to the case.

Keywords. European Union, publications of court decisions, ECLI - European Case Law Identifier, open data, Council conclusions

1. Introduction

On 8 October 2018 the 'Conclusions of the Council and the representatives of the Governments of the Member States meeting within the Council on Best Practices regarding the Online Publication of Court Decisions' (hereinafter also referred to as 'the OPCD Conclusions') were published in the Official Journal of the European Union[2], after they were decided upon in the meeting of the Council on 8 March 2018. In this Chapter various aspects of this document will be discussed. First, Section 2 summarizes the history of this document. Since the document type 'Council conclusions' will be unknown to many, the specificities of this policy instrument will be discussed in Section 3. The sub-

[1] The Author was involved in drafting and negotiating the Conclusions discussed in this Chapter.
[2] OJ C 362/2, 8.10.2018.

stantive contents of the OPCD Conclusions are outlined in Section 4, while in Section 5 the relevance of this document as a soft law instrument is discussed in a broader context.

2. History

Although a legal framework on the publication of court decisions is gradually evolving on the European continent [1], the only official instrument that specifically targeted this issue was Recommendation R(95)11 of the Committee of Ministers of the Council of Europe from 1995[3]. An important incentive to draft that recommendation was the conviction that "*[F]ull knowledge of the jurisprudence of all courts is an essential prerequisite for equitable application of the law*" and that the legal profession as well as the general public should have access to computerized systems for legal research. For several reasons, this recommendation does not cover the actual needs within the EU. First of all, Recommendation R(95)11 only pertains to the access of databases with national court decisions to a national public[4]. Since all Member States of the European Union are Member States of the Council of Europe as well, this in itself would not be problematic if national court decisions would only be a matter of national interest also within the EU. Within the EU though, the Court of Justice already in 1982 decided that national courts have to ascertain that their own opinions on the interpretation of Community law have to correspond with the opinions of the courts of other Member States[5]. This implicit obligation to provide access to national case law collections, was reinforced more explicitly in the resolution of the European Parliament of 9 July 2008 on the role of the national judge in the European judicial system[6]. In other words, from a formal perspective as well as from a substantive angle, the Council of Europe recommendation does not suffice to meet the needs of the European Union. Secondly, although the World Wide Web had been invented by 1995, the Committee could not foresee the implications it would have on the dissemination of court decisions.

Work on the OPCD Conclusions was initiated within 'Building on the European Case Law Identifier' (BO-ECLI), an EU co-funded project[7] that run between October 2015 and October 2017. The project consortium consisted of sixteen organisations from ten EU Member States and was co-ordinated by UBR|KOOP, the Publications Office of the Netherlands. The project had five main objectives. Four of them are not of relevance for the subject of this Chapter, but are mentioned here for the sake of completeness: 1) to start or extend national implementations of ECLI in seven of the participating Member

[3] Council of Europe Committee of Ministers, Recommendation R(95)11 Concerning the selection, processing, presentation and archiving of court decisions in legal information retrieval systems.

[4] On the availability of case law of the European Court of Human Rights (ECHR) itself: Council of Europe Committee of Ministers, Resolution Res(2002)58 On the publication and dissemination of the case law of the European Court of Human Rights. Within the Council of Europe, there is no instrument that stimulates Member States or their courts to provide access to jurisprudence that implement or interpret the European Convention on Human Rights and Fundamental Freedoms or the case law of the ECHR.

[5] Court of Justice, 6 October 1982, 283/81, ECLI:EU:C:1982:335.

[6] (2007/2027(INI)), CELEX:52008IP0352.

[7] Financed from the Justice Programme (2014-2020), call for proposals JUST/2014/JACC/AG/E-JU, action grants to support national or transnational e-Justice projects.

States[8]; 2) to evaluate the current ECLI standard[9] and propose a revised version [3]; 3) to develop an open source software framework for reference parsing [4]; and 4) to promote the ECLI standard among IT and legal professionals[10]. The fifth objective was formulated as: "*To have EU wide policy guidelines on the publication of case law, specifically addressing the issues of selection criteria, data protection and open data*"[11]. The necessary activities to achieve this goal were: "*Analysis of European and national legal and policy frameworks on (various aspects of) case law publication [. . .] and drafting policy guidelines at European level*"[12], while the outputs were defined as: "[A] report, a peer-reviewed paper and (draft) policy guidelines [. . .] for adoption by Council (if politically achievable)"[13].

The analysis was based on a questionnaire and desk research and was delivered in the form of a voluminous report [5] and a summarising academic article [6]. The first part of the report consisted of an aggregated and comparative overview of the state of play regarding existing legal frameworks, actual publishing policies (quantitative as well as regarding quality of access), data protection, open data, citation practice and the implementation of the European Case Law Identifier. The second part contained descriptions regarding those topics for all 28 EU Member States as well as for the European Union itself, the Council of Europe and the European Patent Organization. The report was concluded by small summaries on specific topics, accompanied by 25 recommendations. Since these recommendations were only expressing the personal opinion of the authors, they would not suffice as the 'policy guidelines', since the grant agreement stipulated they should be adopted by the Council. The only type of instrument qualified for this goal are 'Council conclusions'. But since a consortium of public, academic and /or private entities is not in a position to table any item at (the relevant Working parties of) the Council, at least one Member State must initiate such an initiative and take the lead in the drafting and negotiating process. Within the Council Working party on e-Law (e-Justice), this task was taken on by the Netherlands delegation. A first draft of the OPCD Conclusions was discussed in the meeting of 31 January 2017, redrafts were discussed in four consecutive meetings[14] and additional written procedure. The final version was decided upon by Council on 8 March 2018, and published in the Official Journal on 8 November 2018. The first version of the document was quite comprehensive, counting 3.486 words, of which only 851 words survived to the final version. Also, substantively the final version was less ambitious than the initial draft.

Before the content of the OPCD Conclusions is discussed, a clarification is needed on the nature of such 'Council conclusions'.

[8] See for an overview [2].

[9] Council of the European Union, 'Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law', CELEX:52011XG0429(01).

[10] See for an overview of activities http://bo-ecli.eu.

[11] Cited from the grant agreement of BO-ECLI.

[12] *Ibidem.*

[13] *Ibidem.*

[14] 30 March 2017, 4 May 2017, 14 September 2017, 10 October 2017.

3. On the Nature of Council Conclusions

The treaties of the EU are silent about 'Council conclusions', in the Rules of Procedure of the Council[15] they are mentioned but not defined. The nature of this non-legislative instrument is best explained in the 'Comments on the Council's Rules of Procedure':

Conclusions [. . .] are the Council's ordinary means of expression when it is not exercising the powers conferred upon it by the Treaties. In principle, they have the status of purely political commitments or positions with no legal effect. Since they are not legal acts of the Council within the meaning of the Treaties, they are not subject to the procedural rules set out in the legal bases that they might have referred to in view of their substantive content if a decision making power had been exercised in their adoption. Given the lack of a formal adoption procedure in Union law, the Council decides on conclusions by consensus, meaning that a vote is not taken but they are not passed if any Council member opposes them[16].

Nevertheless, even in a non-legislative instrument like conclusions, the Council cannot exceed the powers conferred to it by the Treaties by taking a stance on topics that are not within the Union's areas of activity. However, for political or policy reasons it can be desirable to codify or confirm agreement between Member States on non-EU topics. For such goals also conclusions can be used, but:

Any conclusions referring exclusively to such subjects should be adopted by the representatives of the Member States' governments meeting within (or in the margins of) the Council. If conclusions refer in part to such subjects and no decision has been made to separate Union and non Union matters, the conclusions should be adopted jointly by the Council and the representatives of the Member States' governments meeting within the Council[17].

To avoid (or better: to end) discussions on whether publication of court decisions is a Union matter, the OPCD Conclusions have such hybrid nature. Unlike legal acts, which always have to be published in the Official Journal to obtain legal effect, Council conclusions are only published if this is considered to be necessary or of added value. A decision to publish conclusions is not part of those conclusions themselves, but is a separate decision, taken by the Committee of Permanent Representatives (Coreper) or the Council[18]. Because of their non-legislative nature, conclusions are published in the C-series of the Official Journal, not in the L-series.

4. Substantive Provisions

The OPCD Conclusions contain 17 paragraphs[19], grouped in five sections. The first seven paragraphs – grouped in an untitled introduction – are general in character. Paragraph 1 states the incentives for publication of court decisions (transparency of the judiciary and knowledge of the law), and paragraph 2 mentions the role of the Internet as a new

[15] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l14576.

[16] Council of the EU, *Comments on the Council's Rules of Procedure* (Luxembourg: General Secretariat of the Council, 2016), 67, under a. http://www.consilium.europa.eu/en/documents-publications/publications/2016/council-rules-procedure-comments/.

[17] *Ibidem*, 67, under f.

[18] Art. 17 par. 4 Rules of Procedure of the Council.

[19] The last paragraph is not numbered.

way for judicial authorities to make court decisions available. Paragraph 3 stresses the fact that – as already touched upon above – "*[M]utual knowledge of the legal systems of other Member States is indispensable, especially, but not only, regarding the application of EU law*". Paragraph 4 gives a justification for the existence of the document; after acknowledging that the various interests involved in the online publication of courts decisions require balancing, it states: "*Sharing national best practices can serve as an inspiration for how these interests can be balanced*". Paragraph 5 then starts by stating that these Conclusions describe such best practices, but then continues:

It should be stressed that these best practices lack a mandatory character, do not strive for any kind of harmonisation and should merely be viewed upon as an invitation for reflection. The extent to, and the way in which court decisions are published on the Internet, is up to every individual Member State and/or court to decide.

This 'stating the obvious' demonstrates some reluctancy of Member States to make a (strong) commitment towards those best practices. This caution is also reflected in the terminology of the substantive provisions, with any imperative wording avoided.

Paragraphs 6 and 7 define the scope of the document; paragraph 6 states that 'court decisions' are all types of judicial decisions: "*[U]nder whatever name, rendered by tribunals or courts as defined by national law*". Paragraph 7 defines the meaning of the word 'publication'; it only pertains to the active dissemination to the public at large, it does not cover access to decisions or other courts records under national regimes on access to public documents.

Section I – on selection – only has one paragraph. There are major differences between and within Member States as to the question of selection. In general, a negative selection – publish everything unless evidently not relevant – is followed for the highest jurisdictions, but for first instance and appellate courts the picture is mixed: some courts / Member States do not publish any decision, some make a positive selection (i.e. only publish the interesting cases), while others apply the negative selection for these lower courts as well. Paragraph 8 of the OPCD Conclusions does not give any guidance with regard to the question of selection, but it stresses the need to have selection guidelines in writing and published. Judicial authorities looking for material guidance on selection criteria still have to resort to Recommendation R(95)11.

Section II – paragraphs 9 to 11 – concerns data protection, also a topic on which policies between Member States diverge. Common law systems have a long tradition of publishing full and complete decisions, although gradually a debate evolves on the tension between the underlying open court principle and the risks involved in disclosing personal data in public case law databases[20]. On the other hand, most civil law countries have always followed a strict anonymisation policy for public databases, generally based on the view that personal data of applicants, defendants and witnesses do not serve the primary goals of publication: transparency of the judicial decision process and spreading knowledge about jurisprudential developments. Although the General Data Protection Regulation[21] seems to support the latter view, it leaves ample room for national exceptions. Also here, the OPCD Conclusions do not take a stance, and in paragraph 9 Member States are just called upon to consider the implications of the GDPR for the processing

[20] See, e.g., [7]; [8].

[21] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1, 4.5.2016.

of personal data in court decisions published on the Internet. The words 'if any' in paragraph 10 – "*On choosing a method (if any) for obscuring personal data [. . .]*" – seem to support the view that decisions can be published without any anonymisation, but one could also defend the view that it refers to a situation where the anonymisation is done without any specific predefined method. Paragraph 10 continues: "*[P]reserving the readability and comprehensibility of the text deserves special attention*". Although this is not elaborated, it would imply that solely blanking out personal data or replacing them by 'xxx' is insufficient. Instead, personal data could be replaced by labels ('the defendant', 'license plate', 'witness 13') or fake names, while preferably visualising what has been replaced, e.g. by displaying square brackets around the replacements.

According to article 267 of the Treaty on the Functioning of the European Union national courts are permitted – and if no appeal is possible, they are even obliged – to request for a preliminary ruling from the Court of Justice of the EU if questions arise about the interpretation of the treaties or the validity or interpretation of EU legal acts. As a result of changing publication habits and growing concern about privacy issues, the Rules of Procedure of the Court have been adapted over the years. Currently, article 95 reads:

1. Where anonymity has been granted by the referring court or tribunal, the Court shall respect that anonymity in the proceedings pending before it.
2. At the request of the referring court or tribunal, at the duly reasoned request of a party to the main proceedings or of its own motion, the Court may also, if it considers it necessary, render anonymous one or more persons or entities concerned by the case.

Paragraphs 21 and 22 of the 'Recommendations to national courts and tribunals in relation to the initiation of preliminary ruling proceedings'[22] give further guidance on this. Paragraph 22 reads:

In order to ensure optimal protection of personal data in the Court's handling of the case, [. . .] and the subsequent dissemination [. . .] of the decision closing the proceedings, it is necessary for the referring court or tribunal itself, which alone has full knowledge of the file submitted to the Court, to render anonymous, in its request for a preliminary ruling, the names of natural persons referred to in the request or concerned by the dispute in the main proceedings and to redact any information which could enable them to be identified. Given the increasing use of new information technologies and, in particular, the use of search engines, any anonymisation carried out after the lodging of the request for a preliminary ruling and, a fortiori [. . .] the publication in the Official Journal of the European Union of the notice relating to the case concerned would be devoid of practical purpose.

Since submitting preliminary questions are not daily routine at national courts, there have been many examples where a national court decision containing a question for a preliminary ruling has been published fully anonymised on its own website, while displaying personal information in the publication by the Court in the Official Journal and in the preliminary ruling itself. For this reason, paragraph 11 of the OPCD Conclusions draws specific attention to the topic.

Section III of the Conclusions is about re-use. Paragraph 12 calls upon the Member States to make court decisions that have been published also available for re-use in computer-readable formats. 'Machine-readable format' has been defined in article 2 of

[22] OJ C 257/1, 20.7.2018.

the Directive on Public Sector Information[23] as "*A file format structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure*". Article 5 of this PSI Directive reads:

1. Public sector bodies shall make their documents available in any pre-existing format or language, and, where possible and appropriate, in open and machine-readable format together with their metadata. Both the format and the metadata should, in so far as possible, comply with formal open standards.
2. Paragraph 1 shall not imply an obligation for public sector bodies to create or adapt documents or provide extracts in order to comply with that paragraph where this would involve disproportionate effort, going beyond a simple operation. [. . .]

In paragraph 12 of the OPCD Conclusions an implicit reference to this provision has been made by stating that the recommended good practice is limited: "*[T]o the extent possible given technical or budgetary constraints and given the features of the drafting process*".

No specific formats are mentioned, but in general HTML, Word and PDF are not re-users' favourites. The BO-ECLI report[24] revealed that only in seven Member States (some of the) published court decisions are (also) made available in RDF/XML.

Paragraph 13 makes a comparable recommendation for 'at least formal' metadata, which are to be understood as registrative and objective metadata like the name of the rendering court, the case number and the date of decision or deposit. It can be questioned though whether this constraint has any meaning, since article 5.1 of the PSI directive, cited above, does not make any distinction between types of metadata: if metadata are published, they should be available as open data, whatever their nature.

Paragraph 14 draws attention to the way re-users have access to open data. The BO-ECLI research showed that only nine Member States offer a webservice or FTP-site. In other Member States re-users have to resort to 'screen scraping'. Since this requires re-users to write complex software, potential re-users lacking computer skills or financial resources are not able to obtain the documents. Moreover, traffic on the server will generally be much higher as a result of this screen scraping, which is not always done in the most efficient manner. Therefore, both to the benefit of the suppliers as well as the re-users, paragraph 14 recommends to offer '*adequate download options*'.

Section IV is on improving usability. Paragraph 15 starts with explaining the rationale behind this section: "*Given the vast numbers in which court decisions are published online, not only the bare availability but also the usability of those repositories should be taken into consideration*". Many different options could be considered, but in this paragraph only two – rather obvious – examples are mentioned: search engines and metadata. It is also explicitly stated that the choice of means might depend: "*[O]n the volumes and particularities of the published decisions, actual needs from citizens and the legal community as well as national traditions*".

Paragraph 16 though contains an additional method for improving the usability of databases by tackling the needle-in-the-haystack problem that emerges in voluminous repositories. To separate the wheat from the chaff it is advised: "*To supply for some kind of importance qualification, indicating which, and to which extent decisions are of rele-*

[23] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information as amended by Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013.

[24] See [5], 34.

vance for others than the parties to the case". As was highlighted by the underlying report [5] only few jurisdictions have some kind of a solution to this problem, and hence it can be considered of great importance that this recommendation has survived the weeding exercise that preceded the final version. The text does not give any guidance on how such 'importance qualification' should be implemented, e.g. by human tagging (as done by the ECtHR in the HUDOC database) or by algorithms[25]. Within the BO-ECLI project, as part of its work on 'ECLI 2.0', a proposal was made to formulate a lowest common denominator for expressing such legal importance[10].

Finally, (the unnumbered) paragraph 17 stresses the importance of unequivocal identification and citation and recommends to use the European Case Law Identifier, which was adopted by comparable Council conclusions[26]. Currently, seventeen Member States and three European organisations have adopted ECLI, while others are preparing an implementation[27]. Paragraph 17 also recommends to have ECLI identified decisions indexed by the ECLI Search Engine[28]. Of the 20 jurisdictions that have introduced ECLI in national repositories, currently fifteen are (partly) connected to the Search Engine.

5. The OPCD Conclusions in Perspective

As to their very nature, the OPCD Conclusions are a soft law instrument [11]. Important criteria to judge whether a soft law instrument can considered to be *lex lata* or *lex ferenda*, one has to take into account the intentions of the parties drafting the instrument, the extent to which it reflects current state practice and the relation with other instruments in the field.

With regard to the first criterion – the intentions of the parties – it should be stressed that there was no legal incentive to draft this document; obviously there was a shared opinion that a common vision was needed. Even more, labelling them 'best practices' comes close to a comply-or-explain policy. Also, since publishing Council conclusions in the Official Journal is the exception rather than the rule, the choice to disseminate these best practices reinforces their status.

With regard to the criterion of whether current state practice is reflected, the picture is mixed. The provisions range from definitely reflecting state practice (e.g. the use of (at least basic) search engines and metadata), via gradually evolving state practice (e.g. offering adequate download options for re-users) to not yet established state practice (e.g. tagging published decisions as to their legal importance). However, all these provisions are state practice somewhere, and a clear trend can be noticed.

Regarding the third criterion, the relationship with other instruments, we can witness the gradual development of a legal framework on the publication of court decisions [1]. This framework consists of national legislative and policy instruments as well as soft law from other international organisations. Some authors even argue that these instruments reflect the existence of a human right of public access to legal information. [12]; [13] Apart from the cited Recommendation R(95)11 of the Council of Europe, the ECLI Con-

[25] See, e.g., [9].

[26] Council of the European Union, 'Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law', cit.

[27] State of play is visible on http://bo-ecli.eu.

[28] https://e-justice.europa.eu/content_ecli_search_engine-430-en.do.

clusions and the Resolution of the European Parliament, also the Declaration on Free Access to Law[29] from the Free Access to Law Movement and the 2009 ‘Guiding principles’ of the Hague Conference on International Private law[30] should be mentioned. These latter principles were drafted by the Bureau of the HCCH after consulting specialists from all over the world, and were intended as a foundation for a future global instrument. The principles were reaffirmed and broadened during a joint conference with the European Commission in 2012[31].

One could argue that the OPCD Conclusions are not as detailed or far-reaching as some of the other instruments mentioned. As opposed to those other instruments, the OPCD Conclusions are silent about offering translations, providing access to historic materials or detailed provisions on selection criteria. However, one should not loose sight of the fact that published conclusions of the Council of the EU rank higher than any of the other instruments.

Therefore, these conclusions should be considered a milestone. They could have a serious impact on how courts and court administrations judge their own performance and shape their ambitions, not only within the European Union but all over the world.

References

[1] van Opijnen, M. (2016). Court Decisions on the Internet; Development of a Legal Framework in Europe. *Journal of Law Information and Science, 24*(2), 26, http://www.jlisjournal.org/abstracts/Opijnen.24.2.html.

[2] van Opijnen, M. (2017). Gaining Momentum. How ECLI Boosts Accessibility of Case Law in Europe. *Journal of Open Access to Law, 5*(1), https://ojs.law.cornell.edu/index.php/joal/article/view/57/67.

[3] van Opijnen, M., Palmirani, M., Vitali, F., van den Oever, J. & Agnoloni T. (2017). Towards ECLI 2.0. In Parycek, P. & Edelmann, N. (Eds.). *CeDEM17 International Conference for E-Democracy and Open Government*, Danube University Krems. IEEE, 135-143.

[4] Agnoloni, T., Bacci, L., Peruginelli, G., van Opijnen, M., van den Oever, J., Palmirani, M. et al. (2017). Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links. In Wyner, A. & Casini, G. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 30th Jurix Conference*. IOS Press, 113-118.

[5] van Opijnen, M., Peruginelli, G., Kefali, E. & Palmirani, M. (2017). On-Line Publication of Court Decisions in the EU. Report of the Policy Group of the Project ‘Building on the European Case Law Identifier’.

[6] van Opijnen, M., Peruginelli, G., Kefali, E. & Palmirani, M. (2017). Online Publication of Court Decisions in Europe. *Legal Information Management, 17*(3), 136-145.

[7] Bailey, J. & Burkell, J. (2016). Revisiting the Open Court Principle in an Era of Online Publication: Questioning Presumptive Public Access to Parties’ and Witnesses’ Personal Information. *Ottawa Law Review, 48*, 143.

[8] Winn, P. A. (2004). Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information. *Washington Law Review, 79,* 307.

[9] van Opijnen, M. (2013). A Model for Automated Rating of Case Law. In Verheij, B., Francesoni, E. & Gardner, A. (Eds.). *Proceedings of the 14th International Conference on Artificial Intelligence and Law*. ACM, 140-149.

[29] http://www.fatlm.org/declaration/.

[30] Hague Conference on Private International Law, *Accessing the Content of Foreign Law and the Need for the Development of a Global Instrument in this Area – A Possible Way Ahead*, www.hcch.net/upload/wop/genaff_pd11a2009e.pdf.

[31] Joint Conference of the European Commission and the Hague Conference on Private International Law on Access to Foreign Law in Civil and Commercial Matters, *Conclusions and Recommendations on Access to Foreign Law in Civil and Commercial Matters*, https://www.hcch.net/en/news-archive/details/?varevent=248.

[10] van Opijnen, M. (2016). Towards a Global Importance Indicator for Court Decisions. In Bex, F. & Villata, S. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 29th Jurix Conference*. IOS Press, 155-160.

[11] Terpan, F. (2015). Soft Law in the European Union - The Changing Nature of EU Law. *European Law Journal*, *21*, 68.

[12] Mitee, L. E. (2017). The Right of Public Access to Legal Information: A Proposal for Its Universal Recognition As a Human Right. *German Law Journal, 18*(6), 1429-1496.

[13] Jamar, S. D. (2001). The Human Right of Access to Legal Information: Using Technology to Advance Transparency and the Rule of Law. *Global Jurist Topics, 1*(2).

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190011

Personalised Dissemination of Legal Information

Václav JANEČEK [1]
Faculty of Law, University of Oxford; St. Edmund Hall, Oxford (UK)

Abstract. A proper functioning of any legal system requires people to know the law. Our knowledge of the law, however, depends on how legal information are communicated. Currently, however legal information are communicated rather poorly. We are still missing opportunities that Big Data and algorithms offer in relation to how the law is published, disseminated, and accessed. This Chapter focuses on dissemination of legal information. It argues that we should strive for personalised dissemination. By highlighting and analysing examples from the history of legal publication, it argues that the shift to personalised dissemination of legal information does not pose a threat to the existing legal systems. Instead, it could enhance the overall efficiency and sustainability of our legal communication, increase our knowledge of the law, while reducing the total costs. The Chapter therefore makes a case for a new era in publication and communication of the law – the era of personalised dissemination of legal information.

Keywords. legal information, digitalised legal information, digital law, personalisation, dissemination, communication, legal sources, prehistory, history, hyperhistory, access, printing press, internet, big data analytics

1. Introduction

Modern digital societies are drowning in data that they cannot communicate. The online world is replete with data that are freely accessible, but it is increasingly difficult to retrieve the relevant information from those big datasets. Communication is not only about data transmissions, but also about information production and retrieval. Communication is about exchange of information. Access to data and dissemination of data, must therefore be not confused for access to information and dissemination of information.

Big Data and new analytic algorithms allow us to record, transmit, and newly also process and communicate information. The advertising business has already understood this, as we can witness everyday when we go online and are subject to targeted advertising. It may be true that thanks to new information and communication technologies (ICTs), the swamp of data is getting deeper and deeper, but it is also true that these ICTs are designed so as to not let us drown. ICTs and Big Data combined give us the ability to communicate data more efficiently and not get lost completely.

This includes communication of digitalised legal information. Knowledge of the law in the age of Big Data seems to be withering away precisely because laws are growing

[1] This Chapter builds partly on the Author's previous research [1].

into a big mass of legal data that we cannot communicate. We are sinking in digitalised legal information regardless that we often have free access to data about the law. We are failing at communicating legal information even though legal information are generally much more important than the omnipresent online commercials. Legal professionals may have some ICTs tools to access legal data and retrieve relevant legal information, but the producers of those legal information do not pay much attention to dissemination of any digitalised legal information. In other words, the law-makers struggle to disseminate the law.

The current situation is unhappy because people are expected to know the law and obey it although they cannot efficiently access the relevant legal information and although the law-makers do not bother to communicate the relevant legal information to them. On this front, Big Data and machine learning promises to be a game changer. It promises to facilitate a more efficient and sustainable communication of legal information, particularly as regards its dissemination. We live in a world where you often cannot find the relevant legal information, but where the digitalised legal information could find you. Why, then, do we not yet disseminate legal information in such a personalised manner? Why the law-makers do not communicate the relevant information directly to the addressees of those information, i.e. more efficiently? The standard objection is that we should not try fix things that are not broken, and the current system is not (yet) broken.

The fact that we are not yet drowned does not mean that we should not challenge the current model of publication and dissemination of legal information, provided that its functioning is not designed to prevent us from drowning eventually. My original claim in this Chapter is that there is a case for starting a new era of dissemination of legal information-personalised dissemination of legal information. I propose this solution as a variant that is both feasible and desirable.

The argument unfolds as follows. Section 2 highlights the challenges posed by the growing body of legal data. Section 3 shows why the way in which we have been addressing this challenge so far is and probably will remain unsuccessful. This suggest we should consider unlocking the hidden opportunities presented by new ICTs and Big Data. Section 4 argues we need not be worried to innovate the present models of publication and dissemination of legal information, because we have done it many times before. The history teaches us that revolutionary technologies regularly spur new model of legal publication and that they help increasing efficiency and sustainability of legal communication. Section 5 then discusses some of the apparent costs and benefits of the proposed innovation to conclude that the benefits are likely to outweigh the costs. In a long run, personalised dissemination of legal information therefore seems to be a promising way to make use of algorithms and Big Data as a force for good.

2. The Challenges Posed by the Growing Body of Legal Data

Do you know how many new legal acts or regulations could have applied to you since last year? If you lived in the European Union (EU) in 2018, it would be up to 92 entirely new directly applicable regulations and up to another 132 amendments to the already existing EU law in force[2]. This means that roughly three new regulations were enacted every five days last year. Suffice to add that apart from regulations there are many other

[2] EUR-Lex. *Legal Acts – Statistics (by type)*, 2019, https://perma.cc/2KUD-UG29.

types of legal sources that contain enforceable laws. Now, although these EU laws are freely available via the EUR-Lex online search engine, one has significant doubts about the size of the group out of the EU28's 511.8 million total population[3] that has even the slightest idea about what these regulations regulate. As a lawyer myself, for instance, I know only a fraction of them.

A simple question arises then as to how all the addressees of the EU regulations can comply with such laws if they do not know their content, let alone that those regulations exist. The growing body of legal data presents a significant challenge to our knowledge of the law. The more legal data we produce, the bigger the challenge. The paradox of digitisation in law then is that the more we try to digitise legal data and employ technologies to enhance access to law and justice, the more legal data we produce and efficiently complicate access to law and justice. Big legal data and our ability to obtain relevant legal information from them is one of the pressing issues of the digital societies.

Two obvious ways to meet the challenge are that people will either resign on ever acquiring the relevant legal knowledge or that they will pay massive amounts of money to legal specialists who will communicate the relevant knowledge to them. The first route casts doubts on the role of legal system and we will not discuss this variant. The second setting is, however, also unhappy. It is inefficient and unsustainable because members of society are pushed towards desirable behaviour by lawyers holding an imaginary carrot and stick, instead of them being sustainably educated and informed by relevant legal rules and standards (i.e. legal information). The EU regulations are only rarely discussed publicly – such as in the recent case of the General Data Protection Regulation (GDPR)[4] – and our society at large thus remains legally incompetent, oblivious or even misinformed. This clearly shows a problem with the current model of publication and dissemination of legal information.

3. The Hidden Opportunities of Big Data and Algorithms

The way in which we have addressed the problem so far is insufficient and is lagging behind the opportunities that algorithms and Big Data offer. Take, for instance, the overwhelmingly static model of dissemination of EU legal rules. Those rules are produced in the EU legislative bodies and published in a static journal that is freely available online via the EUR-Lex search engine: the e-OJ (electronic Official Journal of the European Union). The EU citizens who have Internet access and have sufficient personal incentives to learn some legal information can do so freely, provided that they will access actively the static publication platform and provided that they will be able to perform a relatively qualified search task. Alternatively, they can set up an alert system that will feed them with information about every new EU legal document.

In the present model, citizens have free access to legal information, but given the natural human laziness, it seems unlikely that ordinary people, i.e. non-specialists, would be actively searching for legal information in their free time or during their work. Instead, they would be most likely concerned with legal information only when they encounter

[3] Eurostat. *Population on 1 January 2018*, https://perma.cc/NN7P-F4CS.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1, 4.5.2016.

some legal issue (usually presented to them as a legal issue by someone else). At that point, however, the legal issue would probably be so complicated that the non-specialist would still not be able to find the relevant information herself, regardless that her access to law is technically free.

The current system of communication of digitalised legal information is thus clearly underperforming. While access to digitised law is often free, access to legal information (let alone relevant legal information) is expensive. The issue is that non-specialist cannot efficiently access legal information themselves. People may know where they can find the laws and legal data, but they do not know where to look and how to interpret what they see. This seems partly caused by the growing body of freely accessible legal data, partly by the lack of expertise, and partly by the privatised use of expert legal systems by legal professionals. Let us discuss the latest element—the expert systems used by legal professionals.

Legal professionals are increasingly more reliant on new technologies that help them to find the relevant legal data and sometimes event to process the relevant legal information. In the modern era, the ICTs help legal professionals not only to record and transmit legal data, but also to process those data and retrieve legal information from them. These various 'legal information systems' (see, e.g., [2]; [3]) facilitate easier access to law and legal information by providing automated intelligent search engines, assistive predictions about legal information, and computing power that vastly outperforms humans, both specialists and non-specialists. The problem is that while investments in such technologies might be useful for legal professionals, they do not enhance the knowledge of the law by non-specialists. In fact, the increased use of ICTs by legal professionals might lead to gradual propertisation and privatisation of legal knowledge by those experts, because they could eventually represent a limited group of people who can efficiently access legal information. Accordingly, this development would increase the costs barrier when it comes to access to legal information by non-specialists.

Besides, even if we allowed everyone freely to use the expert legal systems to obtain the relevant information and if we built completely free information and AI-based advisory systems[5] [4], this will not address the issue of human laziness. It seem more likely to believe that people would have better knowledge of the law if it was communicated to them not only during legal disputes or via elementary social and educational standards. For the law to serve as an efficient tool to organise society, we should seek actually to increase everyone's knowledge of the law more actively.

In the light of this objective, the hidden opportunities presented by Big Data and analytic algorithms rest, in my view, on the other end of the communication channel. The true challenge is not how we enhance access to legal information, but how we enhance dissemination of legal information. The law-makers should start thinking anew about how to employ new technologies in order to communicate the law's content actively. Today's ICTs allow us to process legal data and legal information in a way that was not only impossible, but also unthinkable a decade ago, and it would thus be a missed opportunity if we only focused on how new technologies can help access the law and legal information.

There are, of course, other ways how to innovate the current model of publication of law [1]. For example, Big Data and new ICTs open up possibilities to personalise the law,

[5] See *Free Access to Law Movement*, http://www.fatlm.org/; *World Legal Information Institute*, http://www.worldlii.org/; *Legal Information Management Journal.*

i.e. to create a specific legal rule applicable to a specific person in a specific situation (e.g. [5]; [6]; [7]; [8]). Such bespoke tailoring of law could be a step forward, but the recent Cambridge Analytica scandal gave us a clear lesson that personalised dissemination of information (the so-called micro-targeting) can be an even more powerful tool because it facilitates efficient communication [9]. My proposition how to meet the challenge is therefore simple. I suggest that the future models of publication of law seek to employ Big Data and new ICTs to personalise dissemination of legal information.

4. The Times They Are a-Changin'

Although the suggested change might seem revolutionary, the history teaches us otherwise. If we look at the history of legal publication models, it seems that new ICTs have often spurred radical changes to how we public, access, and disseminate the law. Let me demonstrate this on some paradigmatic models of publication of legal information. We will see that the progress regarding publication models was often driven by efficiency and sustainability considerations and, therefore, that there is a good reason to adopt personalised dissemination of legal information, provided that such dissemination would increase the overall efficiency and sustainability of how the law is communicated.

The pre-historical era can be our starting point. Since long before people started recording legal information in writing, they used symbolic gestures and pronounced solemn words upon legally important transactions such as when concluding a contract, entering a marriage, etc. in order to mark the importance of the moment and to create at least an impression of the existing legal bond ([10], ch. 3). In such context, it was the witnesses in front of whom these gestures were performed and who thus testified the existence of such legal bond. In this sense, we can say that the witnesses acted as human recorders of relevant legal information, which made it slightly problematic in an environment where people could die easily and where life expectancy was very low. Legal information was thus only a short-term type of information and, on the top of it, it was not very transparent. Similarly, any superior societal authority, i.e. a powerful leader of a social group, who wanted to broadcast his or her own legal information could only rely on this inefficient model. So, for example, when someone got authoritatively punished for trespassing other's assets (e.g. for killing or stealing his life stock), the information about impermissibility of such trespass was only disseminated through stories and vivid experiences of those involved in the execution of the (usually very harsh and painful) physical punishment.

In the context of such unwritten, short-term, and highly volatile system of recording, transmitting and processing of legal information, the rise and implementation of writing as a new technique presented an immense progress. Hence, written records of legal rules and legal information were soon demanded as the new standard. In fact, "a few jurists have believed that paragraphs from the Code of Hammurabi express the obligation imposed by the king to fix in written form contracts (relating to marriage, herding, or tenant farming), or risk having them invalidated" ([10], p. 48). Not only did the invention of writing enable long-lasting access to legal information – e.g. by using clay tablets to record debts, transfers, and property claims in Old Babylonian kingdom ([10], p. 49) – but it also facilitated a more structured and centralised model of publication of legal information. The king could publish all the relevant legal information by a uniform method that was in principle independent of the mortal human ‘recorders’ and ‘interpreters’. One such method to which I will turn below was to set legal rules in stone pillars that were erected in public places.

The dividing line between prehistory and history was never crystal clear because prehistory and history, “prelaw and law, or world of rite and oath on one hand, world of writing on the other” ([10], p. 51) existed in parallel. These two eras were thus inextricably mixed, as for example Babylonian laws from the 20th to 17th century BC demonstrate ([10], p. 51). Accordingly, to speak of a prehistorical model of publication of legal information, we need to focus on its ahistorical features. For the purposes of this Chapter, therefore, the prehistorical model could be best described as a model of shared life experience, a model that could only accommodate and disseminate small amounts of legal information to a small number of people in a limited geographical area and for a short period of time. As such, the prehistorical model was generally inefficient and unsustainable, especially if we take a maturing society ruled by law as our benchmark.

The historical era provides many interesting examples. Some of the oldest written statues, and therefore evidence of some of the oldest historical models of publication of legal information are the famous Babylonian Code of Hammurabi (1754 BC) and Sumerian Code of Ur-Nammu (c. 2100–2050 BC). The Sumerian legal text is still considered the oldest surviving code on the planet ([11], ch. 7). What is typical of these codes is that they were carved into persistent materials such as stone and were stored and displayed at prominent places in the kingdom so that every subject to the king could come and see the binding law (i.e. key legal information) himself or herself. Practically the same model using physically immovable or only hardly moveable materials as a recorder of legal information was used in the Law of Twelve Desks (*Leges Duodecim Tabularum*) in the early Roman Republic (449 BC). Efficiency of this model was based on the fact that people went to public places where these laws were permanently displayed and where everyone could thus learn them. In comparison with the prelaw period, historical legal information could have been more easily preserved over generations and in principle, everyone could have gotten familiar with them. This did not apply only to authoritative rules but also to rules laid down in written contracts.

Still in the historical era, the advancements of ICTs – for instance the invention and implementation of transportable information carriers such as papyrus, parchment, or paper – made it easier to transform the static model of displayed information into a model where the promulgated laws could have been copied (rewritten) and disseminated across the territory in which they were supposed to be binding. Gradually, handwriting on lighter carriers became more widespread and made it possible to disseminate legal information widely. This could be seen as an advantage but also as a problem because legal information was communicated and disseminated rather chaotically, without a clear publication blueprint. The model was becoming less centralised and although common laws could have been transferred from one place to another simply by transporting the relevant legal document, the local laws could have evolved and often also did evolve in substantially different ways.

One way of dealing with this piecemeal publication model was to collect existing historical laws as they developed locally and provide a comprehensive overview of them. This was, for example, the case of *Corpus Iuris Civilis* which was compiled in the 6th century AD subject to order by Eastern Roman Emperor Justinian I, and was then distributed across the Empire as authoritative evidence of existing laws. We can interpret this initiative as an attempt to re-gain control over legal information in the Emperor's territory, and also to monopolise the publication of such information by identifying an authoritative source of that information.

Another interesting model appeared in medieval England, shortly before one such formal source of legal information called Magna Carta Libertatum (1215 AD) was agreed to by King John of England. The English model addressed the problem of decentralisation not by collecting and redistributing common (yet locally distinct) laws, but by implementing a specific system of justice. The efficient dissemination of legal information was facilitated by travelling 'itinerant' judges who dispersed justice according to common laws by going from town to town and hearing disputes of individuals. Paradigmatically, the model thus no longer required the royal subjects to travel to London to find out what the law was. The laws (figuratively speaking) travelled with the judges to the king's subordinates. In a sense, this model was very similar to the one of Eastern Roman Empire, except that it was not an *a priori* written collection of rules, but an *a posteriori* deciding judge who travelled around the country.

Even though it was the official authorities who were responsible for the 'travels' of legal information via the court system and therefore could have maintained control over information dissemination, this model was bound to fail at keeping all published common laws comprehensible. This failure was due to both the increasing amount of judgments, as well as the practical *a posteriori* orientation of the publication model which was not focused on gathering comprehensive information about the common law system, but mainly on applying and disseminating particular legal information as widely as possible. No one could have, at the time, followed the work of every single judge across England, and therefore no one could have known how the common law was evolving. In practice, thus, there was a shortage of access to the increasing body of common law. A provisional solution to this problem was then similar to the model of the Eastern Roman Empire: Sir William Blackstone (18th century AD) collected the existing laws and gathered them into systematically structured volumes entitled Commentaries on the Laws of England. These volumes cleared the way for more principled and comprehensive recording and transmitting of laws. The missing *a priori* piece in the existing *a posteriori* English publication model was thus found.

In the days of Blackstone, one very efficient form of ICT was already in place: the printing press. Historical estimates show that the overall "European book production increased enormously [thanks to advancements in printing technologies] from somewhat more than 12,000 manuscripts per century (or 120 per year) from 500 to 700, to more than one billion books published during the eighteenth century (the peak year in the period 500-1799 is 1790, when more than 20 million copies were printed)" ([12], p. 417).

This made it possible to think anew about how to record and publicise legal information, even though books and printed manuscripts were still regarded as luxury goods ([12], p. 440). The potential for developing a new publication model must have been obvious to anyone in the business of publishing.

The development of legal information and its publication models could have taken multiple routes, but a historical coincidence had it that, at the onset of the 19th century, Europe witnessed an important codification movement resulting in the *Code civil des Français* (*Code Napoléon*), the *Austrian Allgemeines bürgerliches Gesetzbuch* (*ABGB*), etc. The preceding advancements in printing and overall increased literacy in Europe ([13], ch. 4) then supported the idea that laws could be written down in a compact form

– a codification – to be distributed and read widely, much like the Bible. And although positive laws, unlike the Bible, proved to be constantly changing, this historical period made a significant step towards a uniform top-down model of printed publication which was very efficient and which still underpins many of the recent models. The newly gained advantage was that if there were any amendments to the original legal text, the legislator could have only issued the amendment and the addressee could then physically attach this amending piece of information to the designated page in his or her own copy of the codification. Such model allowed for an unprecedentedly wide and efficient dissemination of legal information. The model was, on the face of it, also sustainable.

However, at the same time when the historical publication model started adhering to the idea of codifications, printing was still a low scale business. For instance, the Czech translation of the ABGB from 1812 was never printed in sufficient numbers so that Czech citizens could learn the civil code (ABGB) properly. Evidence suggests that printed translations of the ABGB were scarce and their copies quickly sold out ([14], p. 52).

This was a problem only until mid-19th century, though, when a much more productive rotary printing press appeared. The rotary printing press made things easier but, at the same time, it spurred new troubles. On the one hand, the rotary printing technology in combination with the rise of industrial revolution steeply increased the capacity of printing houses in comparison with those using the outdated flat printing technology. As a result, legislators could now deliver printed legal information to virtually everyone in a relatively short time. The scarcity of printed legal information was no longer an issue. On the other hand, the invention of rotary printing press appeared too late to affect the already existing codification-based as well as printing-press-reliant publication model. It only changed one aspect of this model, namely the method or technology of publication.

One small step for publishers, but one giant leap for legislators. That is how we can describe the change that took place in the second half of the 19th century. The new rotary printing technology, this relatively small advancement of publishers' capacity successfully defeated one important presumption that was present in every legislator's publication model until then. The axiomatic presumption was that laws cannot be changed instantly because communication technologies would not allow for efficient publication of such changes. Towards the end of 19th century, however, this axiom was no longer valid. The development of ICTs (i.e. rotary printing technology) removed factual constraints on how frequently legislative changes and new rules can be made and, as we know from history, our law-making authorities took full advantage of this shift. This was the giant leap for legislators. After this change, new rules and amendments are produced like never before, the rule-making increases and leads to what we may now call a hyperhistorical model of publication of legal information [15]. In the upcoming models we must rely on ICTs not only to record and transmit legal information, but also process it, for otherwise we would not be able to navigate ourselves through this emerging 'hyperlaw' and to manage the enormous amount of legal information.

We have already discussed how digitalised legal information is communicated today. Here we can enrich the argument by several important insights from the history of legal publication models. First, Big Data analytics and self-learning algorithms present a technological challenge that is comparable to the invention of writing or printing press and, therefore, can justify an increased interest in re-thinking the current methods of dissemination of legal information. Second, the innovation regarding dissemination of legal information could be justified if it brings important benefits, including a more efficient and sustainable communication of the relevant legal information. Third, while it can make sense to personalise various aspects of the publication model (including personalisation of the law), the history proves the importance of keeping the law as a system, hence keeping it organised within one model. Historically, this need was manifested in attempts to centralise the publication and dissemination of law. Today, the ICTs allow us to keep authoritative record of legal data even in a decentralised model, but we should not forget that it is a model to serve one system of official state law.

5. Could Personalised Dissemination of Legal Information Be the Right Change?

We can see that the digital challenge compels us to re-think the current publication model of law in order to facilitate more efficient and sustainable communication of legal information. As such, new ICTs and Big Data can be clearly used as a force for good [16]. With an ICTs- and Big Data-powered personalised dissemination of legal information, the relevant law could reach more addressees, and this could lead to an overall increase in our knowledge of the law. Without doubt this amounts a highly valuable good. Besides, the proposed innovation could co-exist with the existing system of authoritative publication of law, which would not undermine the law as a unified system. Yet from a policy perspective, we must ask a more pragmatic question: Would this change be not too costly?

The issue of costs can be dealt with at various levels. We could compare the overall benefits of the envisaged new system with the overall costs of running this system, but then we would probably run into a question whether it is not too costly to have any legal system at all. Given that we have laws and legal systems, we should probably be comparing the costs of the present system of access to legal information with the costs of the suggested personalised dissemination of legal information. We should be comparing these costs in relation to the joint objective of those two alternatives: the costs of efficient and sustainable communication of legal information.

On this level of comparison, the policy question boils down to the question of whether the economic costs of developing and running the suggested system of personalised dissemination of legal information would be lower or higher than money paid by every law-abiding citizen or company for the provision of professional legal services that currently facilitates access legal information. Without empirical data and clear methodology, this question is very hard to answer – especially since the costs of legal services differ significantly across countries and even within the countries. But I doubt many lawyers would bet their money on the current system as the cheaper alternative. In fact, personalised dissemination of legal information could also be cost-efficient for specialists who could save some transaction costs on researching the existing law.

Overall, the suggested method of dissemination of legal information could not only bring more efficient societal organisation and sustainable social cohesion achieved via the increased knowledge of the law, but it could bring down the total costs of access to legal information. Personalised dissemination could close the gap between those who should know the law and those who could access it. The witty idea about this innovative proposition is that legal information would access people, not vice versa. This could not only reduce costs but also overcome the traditional barrier in people's laziness and comfortableness. The new ICTs could communicate the law actively themselves. If we want to use legal information to steer public knowledge, to educate, to unite, and better organise ourselves (eg in the face of global challenges), this could be the way forward.

References

[1] Janeček, V. (2019). Digitalised Legal Information: Towards a New Publication Model. In Öhman, C. & Watson, D. (Eds.). *The 2018 Yearbook of the Digital Ethics Lab*. Springer.

[2] Biasiotti, M. A. & Faro, S. (Eds.) (2011). *From Information to Knowledge: Online Access to Legal Information-Methodologies, Trends and Perspectives*. IOS Press.

[3] Bench-Capon, T. et al. (2012). A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law. *Artificial Intelligence and Law, 20*(3), 215-319.

[4] Greenleaf, G., Mowbray, A. & Chung, P. (2018). Building Sustainable Free Legal Advisory Systems: Experiences from the History of AI & Law. *Computer Law & Security Review, 34*(2), 314-326.

[5] Busch, C. & De Franceschi, A. (2018). Granular Legal Norms: Big Data and the Personalization of Private Law. In Mak, V., Tai, E. T. T. & Berlee, A. (Eds.). *Research Handbook in Data Science and Law*. Edward Elgar Publishing.

[6] Casey, A. J. & Niblett, A. (2016). Self-Driving Laws. *University of Toronto Law Journal, 66*(4), 429-442.

[7] Ben-Shahar, O. & Porat, A. (2016). Personalizing Negligence Law. *New York University Law Review, 91*(3), 627-688.

[8] Porat, A. & Strahilevitz, L. J. (2014). Personalizing Default Rules and Disclosure with Big Data. *Michigan Law Review, 112*, 1417-1478.

[9] BBC (2018). *Cambridge Analytica: Closure 'Will Not Stop Investigation'*, https://www.bbc.co.uk/news/uk-43985186.

[10] Charpin, D. (2010). *Writing, Law, and Kingship in Old Babylonian Mesopotamia* (trans: Todd, J. M.). University of Chicago Press.

[11] Kramer, S. N. (1958). *History Begins at Sumer*. Thames & Hudson.

[12] Buringh, E. & Van Zanden, J. L. (2009). Charting the "Rise of the West": Manuscripts and Printed Books in Europe, a Long-Term Perspective from the Sixth through Eighteenth Centuries. *The Journal of Economic History, 69*(2), 409-445.

[13] Eisenstein, E. L. (2012). *The Printing Revolution in Early Modern Europe*, 2nd ed., Cambridge University Press.

[14] Janeček, V. (2017). *Kritika Právní Odpovědnosti*. Wolters Kluwer.

[15] Floridi, L. (2012). Hyperhistory and the Philosophy of Information Policies. *Philosophy & Technology, 25*(2), 129-131.

[16] Taddeo, M. & Floridi, L. (2018). How AI Can Be a Force for Good. *Science, 361*(6404), 751-752.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190012

Right to Science and Open Access to Legal Knowledge in International and European Law

Gianpaolo Maria RUOTOLO
Università di Foggia (Italy)

Abstract. The Chapter addresses, in an international/EU law perspective, the issue of the dissemination of legal research. The international legal order defines the right to science in the Article 27 of the Universal Declaration of Human Rights; the same right is cited in acts adopted by many international organizations and is included in binding instruments, mainly in the form of the principle of sharing the benefits of scientific research. Affirmed the existence of a right to science in contemporary international law, the Chapter will reconstruct its nature and content: some authors conceive it as an independent right, that deserves an autonomous protection, as it aims at increasing the quality of the life of individuals and collectivities; other scholars build it as an instrument for implementing 'classic' fundamental rights. Among its applications, the one related to the free dissemination of research results, promoted by the Open Access movement, is pivotal, especially with reference to public funded research. In this perspective, the Chapter will mainly focus on three issues: 1) the international law rules on the right to science as legal precursors for open access; 2) the international intellectual property rights regime as a limitation to the right to science and, by the latter, to open access; 3) artificial intelligence, fed by open access, as a means for reconstructing State practice and customary international law.

Keywords. international law, right to science, open access, fundamental rights

1. The Right to Science in the International Legal Order As a Legal Precursor to Open Access (OA): A Basic Legal Framework

The issue of open access to knowledge, with regard to legal knowledge, and more specifically to the works of legal scholarship, can be studied and read in multiple perspectives and from different points of view.

If we look into it by the eyes of an international lawyer, open access is inextricably intertwined with the right to science or, better, it looks as one of the possible ways of implementing the Right to enjoy the benefits of scientific progress and its applications (REBSPA).

In this perspective, the Chapter will mainly focus on three issues: 1) the international law rules on the right to science as legal precursors for open access; 2) the international intellectual property rights regime as a limitation to the right to science and, by the latter, to open access; 3) artificial intelligence, fed by open access, as a means for reconstructing State practice and customary international law.

Let's try to set the normative framework: in international legal order a first enshrining of the right to science is contained in Article 27 of the Universal Declaration of Human Rights (UDHR), whose first paragraph provides that every individual has the right to take part freely in the cultural life of the community, to enjoy the arts and to participate in both scientific progress and its benefits.

Also Article 15, par. 1, of the International Covenant on Economic, Social and Cultural Rights (ICESCR), adopted in 1966 and entered into force ten years later, imposes the States Parties not only to recognize the right of everyone to take part in cultural life, but even, what's more, to enjoy the benefits of scientific progress and its applications, and also to benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author [1].

This right is also often cited in non-binding acts adopted by the United Nations bodies, also dating back: the Charter of Economic Rights and Duties of States, adopted by the General Assembly in 1974[1] [2], contains a right of all the States — but not of individuals — to benefit from both scientific advancement and development in science and technology, and promotes international scientific and technological cooperation between States and the transfer of technology to developing Countries, in order to facilitate access of the latter to modern science and technology (Article 13).

Later on, on 10 November 1975, again the UN General Assembly adopted the more issue-specific Declaration on the Use of Scientific and Technological Progress in the Interests of Peace and for the Benefit of Mankind[2], which provides that all States have to co-operate in the establishment, strengthening and development of the scientific and technological capacity of developing Countries in order to accelerate the realization of the social and economic rights of the peoples of those countries.

The Declaration also invites all States to take any measures to extend the benefits of science and technology to all strata of the population, and hopes that all States shall take the necessary measures, including legislative ones, to ensure that the utilization of scientific and technological achievements promotes the fullest realization of human rights and fundamental freedoms without any discrimination whatsoever on grounds of race, sex, language or religious beliefs.

It also asks States to take effective measures, including legislative ones, to prevent and preclude the utilization of scientific and technological achievements to the detriment of human rights and fundamental freedoms and the dignity of the human person and, whenever necessary, to take action to ensure compliance with legislation guaranteeing human rights and freedoms in the conditions of scientific and technological developments.

Fast forward some twenty years and we reach the Universal Declaration on the Human Genome and Human Rights, adopted by UNESCO's General Conference in 1997 and endorsed by the UN General Assembly in 1998[3], and the International Declaration on Human Genetic Data, adopted by the General Conference of UNESCO on 16 October 2003[4].

[1] A/RES/29/3281, at www.un-documents.net/a29r3281.htm.

[2] digitallibrary.un.org/record/189603/files/A_RES_3384%28XXX%29-EN.pdf.

[3] https://unesdoc.unesco.org/ark:/48223/pf0000122990.

[4] https://unesdoc.unesco.org/ark:/48223/pf0000136112.

The former focuses on risks of the abuse of science and research and less on the sharing of its benefits, while the latter is more centered on the idea of sharing the benefits of science.

Its Article 19 (entitled 'Sharing of benefits'), states that in accordance with domestic law or policy and international agreements, benefits resulting from the use of human genetic data, human proteomic data or biological samples collected for medical and scientific *research should be shared with the society as a whole and the international community*.

On the same issue I have also to remember the Statement on the Right to Enjoy the Benefits of Scientific Progress and its Applications, adopted by UNESCO in Venice in 2009 ('The Venice Statement'), whose para. *iii*) states that advances in information and communication technologies have expanded opportunities for education, freedom of expression and trade, but they have also widened the 'digital gap' [3].

Moving to the European regional level, I recall that the Article 2 of Protocol 1 to the European Convention on Human Rights (ECHR) provides for the right to education ("no person shall be denied the right to education"); the provision, which guarantees a right to access to such education as the State has undertaken to provide, and as regulated by that State, somehow implies the access to the results of scientific research.

Also the Charter of Fundamental Rights of the European Union, provides, in Article 13, the freedom of scientific research and, in Article 14, the right to education (as per Article 2 prot. 1 ECHR).

Outside Europe Article 38 of the Charter of the Organization of American States provides that "the Member States shall extend the benefits of science and technology by encouraging the exchange and utilization of scientific and technical knowledge in accordance with existing treaties and national laws"; Article XIII of the American Declaration on the Rights and Duties of Men states that "everybody has the right to take part in the cultural life of the community, to enjoy the arts, and to participate in the benefits that result from the intellectual progress, especially scientific discoveries" (this is, however, a non binding act of the Organization of American States, so it does not create international obligations on the part of States); Article 14 of the Additional Protocol of San Salvador to the American Convention on Human Rights in the area of economic, social and cultural rights declares that "the States Parties (...) recognize(e) the right of everyone: a. to take part in the cultural and artistic life of the community; b. to enjoy the benefits of scientific and technological progress; c. to the benefit of the moral and material interests deriving from any scientific, literary or artistic production of which he is the author".

Finally, it must be underlined how the Arab Charter of Human Rights, instead, in Article 36 limits to foreseeing that "everyone has the right to participate in cultural life, as well as the right to enjoy literary and artistic aspects", leaving aside the most scientific ones.

2. The Right to Science: Nature and Content

Once affirmed the existence of a right to science in contemporary international law, more ambiguous is the reconstruction of both its nature and content.

Even though, as we've seen, the right to enjoy the benefits of scientific progress and its applications it's not a young one, it is one of the least known and studied human rights ever [4]; [5]; [6].

It has been underlined that "the human rights community has neglected Article 27 of the UDHR and Article 15 of the Covenant" [7]; [8]; [9].

So, despite its aforementioned widespread diffusion in the international legal order sources, the concrete scope of the right to access to scientific knowledge, however, only recently was addressed by the international legal scholarship: it's time that international legal doctrine, looks into the mirror and starts the quest for the right to its own diffusion [10].

About its nature, while some authors conceive it as an independent right, that adds to other fundamental rights and deserves an autonomous protection inasmuch, as it is aimed at increasing the material and spiritual quality of the life of individuals and collectivities, other scholars build it as a mere instrument for the implementation of 'classic' fundamental rights such as culture or health.

Some scholars even see the same right as 'an ideal' one and suggest that it only "constitutes a promising ground (...) in the era of postmodernity" [11], while others theorize that the right to science requires a public good approach to knowledge innovation and diffusion, and that, on this approach it could be the basis to constrain the further expansion of protectionism in international IP law [12].

I think this could be an affordable way to work in order to assure an effective open access to legal knowledge in international law. By now let us keep wiping the dust off the right to enjoy the benefits of scientific progress and give it the attention it deserves [13].

Now, although some of the instruments I mentioned in the first paragraph are not binding, it must be said that, on the one hand, and beyond *ad hoc* disciplines for specific subjects, the international legal order imposes on States that signed the ICESCR – whose number gives the Covenant, at least as regards its substantive rules, a practically universal scope[5] – both the obligation to cooperate to allow the sharing of scientific research results, and the one to adopt positive measures in order to implement the right to science – in its multi faceted aspects – into domestic legal orders.

Let me remind also that the ICESCR has an Optional Protocol which allows its parties to recognize the competence of the Committee on Economic Social and Cultural Rights to consider complaints from individuals, on the model of the Human Rights Committee related to the Covenant on civil and political rights. The Protocol was adopted by the UN General Assembly on 10 December 2008 and opened for signature on 24 September 2009 and, having passed the threshold of required ratifications, it has entered into force on 5 May 2013.

The General Comment n. 21, released in 2009 by the latter Committee, even though focusing on the right of everyone to take part in cultural life (enshrined in Article 15, para. 1 *a*) of the ICESCR)[6] makes clear that all the rights pertaining the cultural/intellectual sphere of the individual are inextricably intertwined.

It clearly states that the latter right is closely related to the other enshrined in the same article, as the right to enjoy the benefits of scientific progress and its applications, the right of everyone to benefit from the protection of moral and material interests resulting from any scientific, literary or artistic production, and the right to freedom, which is an indispensible basis for scientific research and creative activity. And all these rights are also intrinsically linked to the right to education (Articles 13 and 14) and they are also interdependent on other rights enshrined in the Covenant, including the right of all

[5] Nowadays the ICESCR has 169 parties.

[6] Doc. E/C.12/GC/21 of 21 December 2009, at docstore.ohchr.org.

peoples to self-determination (Article 1) and the right to an adequate standard of living (Article 11).

Even though the acts of the Committee are not binding [14], they certainly contribute to the reconstruction of principles and rules related to the issues they deal with, and they influence the practice, even of non-signatory States, that are indeed requested at least to take them into proper account.

Again in order to reconstruct its normative content, in 2012, the Special Rapporteur of the Human Rights Council in the field of cultural rights adopted a Report in which, in addition to linking the right to science to the one to culture and other fundamental rights, tried to define its concrete scope, distinguishing four profiles, constructed, it seems to me, in order to take into account the peculiarities of the right in question in the various international law regimes to which it is applicable.

With this in mind, the Report identifies four distinct expressions of the right to science: the right to access the benefits of science without discrimination; the right to contribute to science and the related freedom of research; the right of individuals and communities to take part in the decision-making processes of States when they concern issues related to science and its impact on the life of the communities themselves, the right to live in a cultural environment conducive to conservation, development and to the spread of science and technology.

It seems to me that at least the first and the fourth points are legal prerequisites for Open Access.

With regard to the subject I am dealing with, moreover, the activities of international organizations is often accompanied by that of non-governmental organizations, individually or in partnership: in particular the International Science Council (ISC) is a non-governmental organization created in 2018 as the result of a merger between the International Council for Science (ICSU) and the International Social Science Council (ISSC) that brings together 40 international scientific Unions and Associations and over 140 national and regional scientific organizations, including Academies and Research Councils. In 2014 the former ICSU released a Report that expressly links "open access to scientific data and literature and the assessment of research by metrics" to the right to share in scientific advancement and its benefits, and to the development of science as a global public good [15].

On these basis, the report asks for the scientific record to be free of financial barriers for any researcher to contribute to, free of financial barriers for any user to access immediately on publication, made available without restriction on reuse for any purpose, subject to proper attribution, quality-assured and published in a timely manner, and archived and made available in perpetuity.

All these goals are now easily reachable by digital technologies.

It remains to be seen to what extent access built this way is fully compatible with international standards that protect intellectual property rights, and whether these are really in conflict with the right to science and really represent an impassable limit to open access [16]; [17]; [18].

3. Open Access and Intellectual Property: The Berne Three Step Test As a Model

The at least potential conflict between the right to science in its open access embodying and the international instruments of protection of intellectual property rights is well known.

But we should also remember that in some sectors of IP law weird things happen: let us think to the case of the so called ‘cover songs’, wherein the copyright laws contain provisions for the compulsory licensing of musical compositions, that enable every musician to play his own version of an existing tune (after its first commercial exploitation), so that the composer of the song cannot prevent it, and is just entitled to a reasonable royalty.

So in IP law not everything is always what it may seem at a first glance.

If we switch back to the legal knowledge, there is to say that rules that are easily communicated and understood will be applied frictionless and efficiently (one may think to traffic laws, which are signaled with roadside signs) and that the more complex laws are, the less they are accessible, even though they are uncopyrighted and uncopyrightable, and for this they require the interpretive skills of a lawyer [19].

And the latter kind of legal information is copyrightable and indeed treated in very proprietary ways; books, papers, digests, and even annotated versions of laws and statutes – what the legal scholarship is – are mostly copyrighted and accessible through specialized databases.

And we have also to bear in mind that, as for Article 38 of the Statute of the International Court of Justice, in the international legal order legal scholarship is a source of law, even though a subsidiary one ("...the teachings of the most highly qualified publicists of the various nations, as subsidiary means for the determination of rules of law").

It is also often believed that the international patent system represents a limit to the circulation of information, culture and therefore, ultimately, to the right to science; but it must be also remembered that both the approved and the rejected patents, with the related instances and attached documentation, are freely accessible through public databases: the patent mechanism could be seen as some kind of a first model of ‘pseudo’ open access.

In the perspective of knowledge, indeed, what the patent prevents is not so much the knowledge of the patented invention itself, as its exploitation. Even though for some categories of goods, mere knowledge is not sufficient to fill the gap (for example, knowing the existence of a patented pharmaceutical product does not solve the problem of the access to the same by those who, while needing it, can’t afford it) there is to say, that such a knowledge could cause the interested individuals, for instance, to lobby with the competent bodies to obtain it as a benefit or by compulsory licensing.

Obviously this could not be *sic and simpliciter* applied to law.

And neither, of course, I am saying that I am thinking of the patentability of legal works as a solution to the problem of the access to knowledge: I am just trying to see things differently and switch paradigm.

Now, as any limitations on copyright must accord with international obligations, I have to remember that all the main international law instruments in IP (Berne Convention, TRIPs, WIPO Copyright Treaty and Performances and Phonograms Treaty), and even most of the EU pertaining rules, contemplate exceptions to exclusive rights, mostly built on the so called ‘Berne three-steps test’ [20].

The three-step test, first established in 1967 pertaining the exclusive right of reproduction under Article 9(2) of the Berne Convention, holds that States may proscribe copyright limitations only "1) in certain special cases, 2) provided that such reproduction does not conflict with a normal exploitation of the work and 3) does not unreasonably prejudice the legitimate interests of the author".

In this context it is impossible for me to carry out a thorough examination of the elements of this text; in any case, it is sufficient to point out that the domestic and international case law has on several occasions clarified them.

For a guide on the way the test can be performed I will refer to World Trade Organization (WTO) practice.

As regards the requirements referred to in no. 1), the Panel appointed for the *US copyright* case has made it clear that copyright restrictions that are undetermined or unspecified are prohibited (*certainty* requirement). As for the *specialty* requirement, always referred to in n. 1), the same panel has identified both a quantitative and a qualitative element: the former requires that the exception can be applied only to a limited number of cases, and the latter that it must be applied to achieve a well-defined policy objective[7].

The second step of the test provides that any limitation or exception must not conflict with a normal exploitation of the work; the same WTO panel determined that an exception or limitation violates this step "if uses, that in principle are covered by [copyright] but exempted under the exception or limitation, enter into economic competition with the ways that right holders normally extract economic value from that right to the work (i.e., the copyright) and thereby deprive them of significant or tangible commercial gains".

To pass the third, final, step, any copyright limitation or exception must "not unreasonably prejudice the legitimate interests of the author": the WTO panel on *US Copyright* dispute made clear that "the notion of 'interests' is not necessarily limited to actual or potential economic advantage or detriment (...); it has also the connotation of legitimacy from a more normative perspective, in the context of calling for the protection of interests that are justifiable in the light of the objectives that underlie the protection of exclusive rights"[8].

On these basis, a fair application of the three-step test may lead to a wider legitimacy of OA in the international legal order as a means of applying the internationally protected right to science.

Moreover, this would also make it possible to compose a fragmentation of the international legal order, by resorting to the systemic interpretation of Article 31, par. 3, lett. *c)* of the Vienna Convention on the Law of Treaties, according to which, as is known, in the interpretation of a treaty, any relevant rule of international law applicable to relations between the parties must be taken into account [21].

And to similar results could lead, albeit in a more nuanced manner, the fair use exception of US law, that uses a four factor test[9] [22].

[7] See Panel Report WT/DS160/R, *United States-Section 110(5) of the US Copyright Act*, WT/DS160/R, 15 June 2000. On the Berne test see also, again in WTO context, Panel Report WT/DS1 14/R, *Canada-Patent Protection of Pharmaceutical Products*, 17th March 2000 and Panel Report WT/DS174/R, *European Communities-Protection of Trademarks and Geographical Indications for Agricultural Products and Foodstuffs*, 15th March 2005, all available at www.wto.org.

[8] At 11 6.223-6.224.

[9] 17 U.S.C. § 107: Limitations on exclusive rights: Fair use (West 2011). "the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means (...), for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include – (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work".

4. A Data-driven Right To Know International Law

In a paper published some years ago [23] I tried to investigate the way the widespread diffusion of the Web impacted on the way international legal order operates.

Then I decided to adopt a, so to speak, heuristic approach – an approach that is applied with greater frequency in sciences other than legal, such as social or information sciences: I did it in order to identify major trend lines when, as in that case, timely analysis of all the data relating to the observed phenomenon is impractical. I mean that it was then impossible, for me at least, to collect and study *all* the practice pertaining the issue I was going to investigate. And heuristics could then help me to facilitate the access to new empirical findings by means of a process that, to solve a given problem, relied on the contingent state of factual circumstances.

Whenever such a knowledge is impossible for humans, and this happens more and more in the case of complex systems such as legal systems, an overall assessment in the light of a partial documentation, by this approach, is possible, and it indeed could better capture trend lines.

But this perspective can be outmatched by OA legal Big Data filtered by AI algorithms, linked data – a method of publishing structured data so that it can be interlinked and become more useful through semantic queries – and even micro-task crowdsourcing (Amazon's Mechanical Turk, CloudCrowd, CrowdFlower), as this can help us "to look at all the available data rather than subsamples thereof" [24] and let us know much deeper our issues.

In short, open access to legal knowledge, guaranteed through the computer tools now commonly used, could allow appropriately trained expert systems, and in particular artificial intelligence mechanisms, to put together all the international practice relating to a given topic and simplify, consequently, the construction of the material element of customary international law. This would not only allow the international law scholarship to globally examine itself, in that game of mirrors to which we referred at the beginning of our discourse, but could also lead to increase the knowledge of the rules of international law by their addressees and, consequently, to improve their effectiveness.

I would close reminding that if the knowledge international law refers to is a global common, the *Declaration on Free Access to Law* adopted by our predecessors in the fourth edition of LVI Conference (in 2002 in Montreal) qualifies the legal information as a common heritage of mankind[10], and clarifies how its massive diffusion can contribute to the respect of the rule of law.

References

[1] Green, M. (2000). *Drafting History of Article 15(1)(c) of the International Covenant on Economic, Social and Cultural Rights*. UN Doc. E/C.12/2000/15, 9 October 2000.

[2] Chatterjee, S. K. (1991). The Charter of Economic Rights And Duties of States: An Evaluation After 15 Years. *International and Comparative Law Quarterly, 40*(3), 669-684.

[3] Müller, A. (2010). Remarks on the Venice Statement on the Right to Enjoy the Benefits of Scientific Progress and Its Applications (Article 15(1)(b) ICESCR). *Human Rights Law Review 10*(4), 765-784.

[10] It is appropriate to highlight how the doctrine has already qualified knowledge as a global public good to which, consequently, universal access should be guaranteed: see [12], p. 156.

[4] Donders, Y. (2011). The Right to Enjoy the Benefits of Scientific Progress: In Search of State Obligations in Relation to Health. *Medicine, Health Care and Philosophy, 14*(4), 371-381, www.springerlink.com/content/xk66t30m93122853/fulltext.pdf.

[5] Chapman, A. (2009). Towards an Understanding of the Right to Enjoy the Benefits of Scientific Progress and Its Applications. *Journal of Human Rights, 8*(1), 1-36.

[6] Vadi, V. (2008). Sapere Aude! Access to Knowledge As a Human Right and a Key Instrument of Development. *International Journal of Communications Law and Policy, 12*, 346-368.

[7] Chapman, A. R. (2001). Approaching Intellectual Property As a Human Right: Obligations Related to Article 15 (1) (c). *Copyright Bulletin, 35*(3), 4-36.

[8] Schabas, W. A. (2007). Study of the Right to Enjoy the Benefits of Scientific and Technological Progress and Its Applications. In Donders, Y. & Volodin, V. (Eds.). *Human Rights in Education, Science and Culture*. Ashgate Publishing, 273-308.

[9] Riedel, E. (2012). Sleeping Beauty or Let Sleeping Dogs Lie? The Right of Everyone to Enjoy the Benefits of Scientific Progress and Its Applications (REBSPA). In Hestermeyer, H.P., König, D., Matz-Lück, N. et al. (Eds.). *Coexistence, Cooperation and Solidarity*. Brill Nijhoff, 503-519.

[10] Caso, R. (2009). Open Access to Legal Scholarship and Copyright Rules: A Law and Technology Perspective. In Peruginelli, G. & Ragona, M. (Eds.). *Law Via The Internet. Free Access Quality of Information Effectiveness of Rights*. European Press Academic Publishing, 92.

[11] Andersen Nawrot, A. M. (2014). *The Utopian Human Right to Science and Culture: Toward the Philosophy of Excendence in the Postmodern Society*. Routledge.

[12] Shaver, L. (2010). The Right to Science and Culture. *Wisconsin Law Review, 1*, 121-184.

[13] Donders, Y. (2007). *The Right to Enjoy the Benefits of Scientific Progress and Its Applications*. Speech on the Occasion of Human Rights Day, UNESCO Headquarters (Paris, 10 December 2007), https://unesdoc.unesco.org/ark:/48223/pf0000158691.

[14] Keller, H. & Grover, L. (2012). General Comments of the Human Rights Committee and Their Legitimacy. In Keller, H. & Ulfstein, G. (Eds.). *UN Human Rights Treaty Bodies: Law and Legitimacy (Studies on Human Rights Conventions*, 116-198. Cambridge University Press.

[15] Dalrymple, D. (2003). Scientific Knowledge As a Global Public Good: Contributions to Innovation and the Economy. In *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*, 35-49. National Academies Press.

[16] Anderson, R. D. & Wager, H. (2006). Human Rights, Development, and the WTO: The Cases of Intellectual Property and Competition Policy. *Journal of International Economic Law, 9*(3), 707-747.

[17] Chapman, A. R. (1999). A Human Rights Perspective on Intellectual Property, Scientific Progress and Access to the Benefits of Science. In WIPO/OHCHR. *Intellectual Property and Human Rights*. Panel Discussion to Commemorate the 50th Anniversary of the Universal Declaration of Human Rights (Geneva, 9 November 1998), 127-168.

[18] Plomer, A. (2013). The Human Rights Paradox: Intellectual Property Rights and Rights to Access to Science. *Human Rights Quarterly*, 143-175.

[19] Bartow, A. (2006). Open Access, Law, Knowledge, Copyrights, Dominance and Subordination. *Lewis & Clark Law Review, 10*, 869.

[20] Iuliano, J. (2011). Is Legal File Sharing Legal? An Analysis of the Berne Three-step Test. *Virginia Journal of Law & Technology, 16*, 464.

[21] McLachlan, C. (2005). The Principle of Systemic Integration and Article 31(3)(c) of the Vienna Convention. *International and Comparative Law Quarterly, 54*(2), 279-320.

[22] Nguyen, N. A. (2010). Not All Textbooks are Created Equal: Copyright, Fair Use, and Open Access in the Open College Textbook Act of 2010. *DePaul Journal of Art, Technology & Intellectual Property Law, 21*, 105.

[23] Ruotolo, G. M. (2013). The Impact of the Internet on International Law: Nomos without Earth? *Informatica e diritto*, 2, 7-18.

[24] Alschner, W. Pauwelyn, J. & Puig, S. (2017). The Data-Driven Future of International Economic Law. *Journal of International Economic Law, 20*(2), 217-231.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190013

Open Science, Open Doctrine, How to Share Knowledge?

Marie FARGE [a] and Jean GASNAULT [b]

[a] *CNRS, Institut National des Sciences Mathématiques et de leurs Interactions - INSMI; Ecole Normale Supérieure - ENS; Committee for the Accessibility of Publications in Sciences and Humanities - CAPSH (France)*

[b] *La Loi des Ours; Université Paris 1, DU Droit et Informatique (France)*

Abstract. The evolution of Open Science in France is almost completely the result of constant friction with the business models that drive major international publishing houses, where each party has adapted to developments introduced by the other, but also of practical steps taken to ensure that shared documents are efficiently collected and made accessible. This Chapter will provide several examples of the development of Open Science in France, such as the platform http://dissem.in. How Open Science principles are effectively implemented in the area of legal knowledge in France? What can be done to encourage law scholars to publish their work on a single common platform? And which platform should that be? Should it be improved, and, if so, in what way? Will dialogue resolve conflicts and pave the way for Open Science in a viable economic context?

Keywords. open science, legal doctrine, open access

1. Introduction

The Free Access to Law Movement (FALM)[1] has a long history of promoting free access to legislation and case law, especially in common law countries. The European Union, its Member States, several OECD Member States and the signatory countries of the Hague Conventions have similar objectives. The conclusions and recommendations published in 2012 by the Hague Conference on Private International Law (HCCH) and the European Commission prove this. Where are we with regard to the free and open sharing of scientific knowledge in law? Have practical measures been implemented to improve the sharing of legal doctrine and commentary?

A number of comparative approaches can be adopted to identify and measure recent developments in the sharing of legal science. In this context, an electronic debate on the challenges of open access for legal researchers took place between 2016 and 2017 on the ‘BlogdroitEuropéen’. The Chapter will present the results from an interdisciplinary perspective and will attempt to determine whether French lawyers have been in line with the approach followed by their colleagues in the sciences.

[1] http://www.fatlm.org/.

While the prevailing working methods of scientists and lawyers correspond to different academic cultures (e.g., the ways to become a professor of mathematics or a professor of law cannot be compared), there are undoubtedly similarities for publishing and peer-reviewing. In France, social sciences and exact sciences academic worlds have started to engage with one another in a number of settings (symposiums, non-profit associations, working groups), with the common goal of harnessing the benefits of Open Science. They are striving to make scientific publications easier to find, and harmonize the way they are described and indexed. By mobilizing and joining forces, their goal is to disseminate the results of scientific research as widely as possible. Although the two communities followed different paths, they reached the same conclusion: research progress inevitably requires sharing, which entails no losses, only gains[2].

A number of principles have been established over the past 15 years. The Open Access Initiative, announced in Budapest in 2002 and developed in 2003 in the Berlin Declaration, was a pioneering project. Gradually, legal norms and standards emerged, starting with a recommendation from the European Union, which has become the fully-fledged program Horizon 2020. The objective of all these actions is to promote the widest possible access to knowledge, with the clear risk of conflict with existing economic paradigms.

In France, Article 30 of the 2016 *Law for a Digital Republic*[3] marks the first significant victory for the advocates of digital commons[4]. Unfortunately, this measure was poorly applied and it has received very little publicity from the government and any from the knowledge market players. Thus, although clear principles and laws have been established, their translation into national legislation has proved difficult. Could this be interpreted as a symbol of Open Science? We can only hope that the situation will change. The French government has reiterated its commitment to the development of Open Science in the 'Etalab Action Plan' for 2018-2020[5].

This Chapter will address the two following topics[6].

Section 2 focuses on the development of Open Science, which depends on the international confrontation between the business model of the few major publishers currently dominating scientific publishing and the need of researchers to peer-review, publish and access scientific articles, as easily, widely and efficiently as possible. We will describe the current situation in the world, and then we will focus on France with some specific examples.

Section 3 is devoted to the effective implementation of the principles of Open Science in the field of legal knowledge in France.

[2] The Open Science Committee set up by the French Ministry of Research mobilises actors in the scientific world, regardless of their discipline https://www.ouvrirlascience.fr/presentation-du-comite/.

[3] *Loi n^o 2016-1321 du 7 octobre 2016 pour une République numérique*, https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id.

[4] Lionel Maurel's blog post, SILEX, 31 October 2016, *Open Access, quelles incidences de la loi «République numérique»?* https://scinfolex.com/2016/10/31/open-access-quelles-incidences-de-la-loi-republique-numerique/.

[5] Etalab (French task force for Open Data). France's Open Government action plan 2018-2020, *Engagement 14: Construire un écosystème de la science ouverte*, https://gouvernement-ouvert.etalab.gouv.fr/pgo-concertation/topic/5a1bfc1b498edd6b29cb10d4.

[6] Marie Farge is the Author of Section 2. Jean Gasnault is the Author of Section 3.

2. Open Science

2.1. The Electronic Publishing Revolution

The reproducibility of published results is the backbone of scientific research. Objectivity is crucial for science and requires that observations, experiments and theories be checked independently of their authors before being accepted for publication. Indeed, a result to be recognized as scientific must be presented and explained in an article which has been reviewed and accepted by peers, i.e., researchers able to understand, verify and, if necessary, correct it. It is only after successful peer review that a new result can be published and belongs to scientific knowledge. Consequently, the set of all scientific publications is the common heritage that researchers have collectively built over centuries, and are constantly developing. Given the constructive and universal nature of science, any researcher should have access, as early and easily as possible, to all scientific publications. Unfortunately, this is not the case today, as most peer-reviewed journals belong to a few major publishers, who keep scientific articles behind pay-walls. Since all over the world the majority of research programs are supported by public funds financed by taxpayers, not only researchers, but everyone from everywhere should have access to scientific publications.

Before the advent of electronic publishing and of the Web, researchers had never criticized the business model of scientific publishing, where journals were paid by subscription, since there were no other ways for their articles to be disseminated and read. In those years, researchers were sending handwritten manuscripts to be peer-reviewed by researchers who are experts of the scientific domain covered by the journal. Publishers were printing houses in charge of typesetting, printing and selling the journals to libraries. Today, the era of paper publishing is over and replaced by the era of electronic publishing. Indeed, recent articles, as well as older ones that have been digitalized, are exchanged electronically via the Web. Even for journals that are still printed on paper, their production is made electronically. Moreover, most readers download articles from the Web and print them only if needed. This technological revolution has allowed publishers to drastically reduce their costs, and researchers to typeset their articles themselves, while both peer-reviewing and publishing are made online via electronic platforms. Under these conditions it is unfair that publishers still maintain the business model of 'paper publishing' and make skyrocketed profits (up to 40%, i.e., twice those of Google or Apple) using 'electronic publishing'. The explanation for such impressive profits is simple: the investments for producing scientific results and articles are publicly funded, while the ownership of scientific journals and corresponding profits are private.

Foreseeing the opportunity of electronic publishing, a few major companies (Elsevier, Springer Nature, Wiley-Blackwell, etc.) have succeeded in buying most scientific publishers and now own most of the journals researchers need, for peer-reviewing, publishing and sharing their results. They therefore control the market and impose pay-walls to access to articles, which deprive researchers from some of their public funding. Moreover, few dominant publishers have managed to get the vertical control of research since they own, not only scientific journals, but also: the platforms which researchers use for peer-reviewing (e.g., Evise-Elsevier) and for publishing (e.g., ScienceDirect-Elsevier); the platforms which librarians use for bibliometry (e.g., Scopus-Elsevier); the platforms which managers use to evaluate researchers (e.g., SciVal-Elsevier).

Indeed, quantitative indicators designed, controlled and owned by a very small number of publishers hand them control over research policy, which they did not have in the past. Researchers deplore the fact that today these major publishers control their research activity by selling, not only their articles, but also the tools to evaluate their careers and sort their applications for research contracts.

2.2. *Open Science in the World*

Open Science means free access to any research output (peer-reviewed articles, conference proceedings, data, software, blogs, ...) and to the metadata describing them. The objective of Open Science is to help researchers from all over the world to collaborate, interact, share resources and disseminate results, as freely, rapidly and efficiently as possible, by taking advantage of electronic publishing. This movement is gaining strength worldwide, thanks to more than thirty years of efforts to advocate it and develop the necessary tools (infrastructures, software, etc.). Unfortunately, the majority of scientific journals which researchers use to peer-review and share their results are still owned by a few major publishers which control scientific publishing worldwide. Moreover, publishers also own most of the peer-reviewed articles: their text, data tables, figures and any other additional material (e.g., codes) that researchers are obliged to give them for free, as soon as their articles have been accepted for publication. The ownership of a journal is forever, unless the publisher ends its publication; if a journal is sold or given away, the ownership is transferred to the new publisher. The ownership of articles lasts up to seventy years after the death of the last co-author; publishers thus own and control the access to the last one hundred years of scientific research, at least! It is also unfair that the prestige of a scientific journal depends on the commercial strength of a few major publishers and of their practice of 'bundling' (e.g., Elsevier negotiates only one contract to sell the access to the 3,800 journals of its electronic platform 'Science Direct'). The scientific quality of a journal should rather depend on the expertise and the dedication of its editors and the referees they choose (i.e., researchers who volunteer their time free of charge to review the articles submitted to the journal).

Here are some of the key steps that contributed to the development of Open Science.

In 1974, Donald Knuth, professor of computer science at Stanford University, designed a typesetting software for text and mathematical formulae, called 'TeX' (from the Greek word for art, skill, craft), and published it in open-source for anyone to use. In 1983, Leslie Lamport, mathematician and computer scientist at the Stanford Research Institute, enhanced TeX by incorporating a set of macros that separated content and style in the document. This made TeX easier to use and it became LaTeX[7], which is today the standard format for articles in physics, mathematics and computer sciences. It has also been adopted by researchers in other disciplines, such as economics and the history of science, as it allows the text to be modified independently of the layout.

In 1989, the physicist Joanne Cohn, from the Institute of Advanced Studies in Princeton, created an e-mailing list for sharing preprints of string theory, which was made possible thanks to the low-bandwidth of the TeX format which theoretical physicists were beginning to use.

In 1990, the physicist and computer scientist Tim Bernes-Lee created at the European Center for Nuclear Research (CERN) the open protocol Hyper Text Transfer Protocol (http) and, in order that it could be adopted by anyone, he decided not to patent it. This was the birth of the World Wide Web (WWW).

[7] https://www.latex-project.org.

In 1991, the physicist Paul Ginsparg, from the Los Alamos National Laboratory (LANL), automated the e-mailing list of Joanne Cohn by using the File Transfer Protocol (ftp) and the http protocol of the Web in 1993. This was the birth of the open repository arXiv[8], which later was moved to Cornell University in 1999. Today, the majority of articles in mathematics, computer sciences and physics are deposited by their author on arXiv as soon as they are submitted to a peer-reviewed journal, or even before. Researchers are keen to follow which new articles of their research domain have been deposited, since the platform arXiv informs them by emails and RSS feeds.

In 1994, the economist Michael Jensen (Harvard University, USA) created the Open Platform Social Sciences Research Network (SSRN) to share preprints for social sciences, the largest open repository in 2013, but Elsevier bought it in 2016 and researchers lost control of this research tool.

In 1998, the professor of learning sciences and technology design John Willinsky (Stanford University, USA, and Simon Fraser University, Canada) released the open-source software Open Journal System (OJS)[9] to manage editing and peer-reviewing, which is used today by more than 10,000 academic journals.

In 1999, the biochemist Rogerio Meneghini and the information scientist Abel Packer, both from the Federal University of Sao Paulo in Brazil, designed the platform Scientific Electronic Library Online (SciELO)[10] to increase visibility and access to research publications, especially for countries which cannot afford paying the ever increasing subscription costs required by academic journals. It is a network of 14 countries, from Latin America, Caribbean, Portugal, Spain and South Africa, which develop a common methodology for the preparation, storage, dissemination and evaluation of scientific articles of all disciplines. SciELO enables the electronic publication of more than 1,000 peer-reviewed journals which are selected for their scientific quality, the organization of searchable bibliographical and full text databases, the preservation of electronic archives and the production of statistical indicators of the scientific literature usage and impact, including journal evaluation criteria. SciELO is funded by several public institutions from Latin America and Spain.

In 2000, a few researchers around Harold Warmus (Nobel Prize winner and former director of the National Institutes of Health, a federal agency of United States) launched an online petition which called for all scientists to pledge that from September 2001 they would discontinue submission of articles to journals that did not make the full text of their articles available to all, free and unfettered, either immediately or after a delay of no more than 6 months. This led to the creation of PubMed Central (PMC)[11], an open repository that archives articles published in biomedical and life sciences journals, and to the creation of the Public Library of Science (PLOS)[12], a non-profit organization which publishes in open access seven biology and medicine academic journals.

In 2002, 16 researchers and librarians, from Europe, Canada and United States, met in Budapest and launched the already mentioned Budapest Open Access Initiative calling for Open Access to all scientific publications which should be considered as a 'public good'. They explained that "Old tradition and new technology have converged to make

[8] https://arxiv.org.

[9] https://pkp.sfu.ca/ojs.

[10] http://www.scielo.br.

[11] https://www.ncbi.nlm.nih.gov/pmc.

[12] https://www.plos.org.

possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the Internet. The public good they make possible is the worldwide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge. For various reasons, this kind of free and unrestricted online availability, which we will call open access, has so far been limited to small portions of the journal literature".

In 2003, the presidents of 19 national research institutions from Germany, France, Italy, Spain and Hungary published the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[13], which states that: "Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material. [...] The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship. [...] Our organizations are interested in the further promotion of the new open access paradigm to gain the most benefit for science and society".

After nearly twenty years of fierce resistance, the major publishers owning most of the peer-reviewed journals have now accepted to publish some articles in open access, but under the condition that authors, or their institution, pay them Article Processing Charges (APCs). Publishers gave the fancy name 'Gold Open Access' to this business model, where they still own the peer-reviewed journals and fix the price of APCs in a way that preserves their huge profit margins (i.e., up to 5,000 € per article, or even higher). The impressive lobbying that publishers do, either directly, or via associations, e.g., STM (Science, Technology and Medicine) and AAP (Association of American Publishers), in Brussels, Washington, Beijing, London and elsewhere, is extremely efficient since they are wealthier than the largest research institutions in the world; e.g., in 2017 the turnover of Reed-Elsevier was 8.4 billion € while the budget of the French National Center for Scientific Research (CNRS), the largest in Europe, was 3.3 billion €. In 2012 a group of mathematicians launched the movement The Cost of Knowledge[14], called to boycott of Elsevier and succeeded in stopping the Research Work Act, a bill of the American Congress that Elsevier had lobbied for. The associations STM and AAP together with Elsevier then redirected their lobbying towards Europe. In July 2012 the Gold Open Access model became mandatory in the United Kingdom for articles whose research has been supported by the UK Research Council or the Wellcome Trust. The same month the European Commission also recommended Open Access, but offered researchers the choice between two business models, defined as follows: "Gold Open Access (open access publishing): payment of publication costs is shifted from readers (via subscriptions)

[13] https://openaccess.mpg.de/67605/berlin_declaration_engl.pdf.

[14] http://thecostofknowledge.com.

to authors. These costs are usually borne by the university or research institute to which the researcher is affiliated, or by the funding agency supporting the research. Green Open Access (self-archiving): the published article or the final peer-reviewed manuscript is archived by the researcher in an online repository before, after or alongside its publication. Access to this article is often delayed (embargo period) at the request of the publisher so that subscribers retain an added benefit"[15].

To resist the lobbying exerted by a few major publishers who try to control electronic publishing and Open Access, the researchers from The Cost of Knowledge movement proposed in June 2012 a third model called 'Diamond Open Access' (a terminology inspired from the Diamond Sutra which is the oldest printed document made in 868 in China)[16]. This model is characterized by the fact that readers and authors should not pay to read or publish research articles. It is based on the following principles:

- Authors keep their copyrights and attach to their article a Creative Commons License CC-BY (allowing everyone to publish, use or translate their article while only requiring the attribution of the paper to the authors).
- Editorial boards are legal entities which own peer-reviewed journals (i.e., its title and all its assets), whose members are researchers who take the responsibility of peer-reviewing without being paid (since it is part of their academic duty for which they get a salary).
- Publishers are no longer the journal's owners, but become service providers under contract with the journal's editorial board.

In order to reduce the journal's cost, peer-reviewing and publishing can be automated using software, as commercial publishers do for the journals they own. But there is an essential difference with the Diamond Open Access model because researchers use free open-source software that they have developed to match their needs, e.g., LaTeX and OJS developed by the Public Knowledge Project (PKP)[17]. If a Diamond Open Access journal is recognized to be useful to its scientific community, and as long as its editorial board can prove good peer-reviewing practice, it could be published for free using the services of a publishing platform, which is a publicly-owned and publicly-funded infrastructure, designed to service a very large number of journals from different fields. The dissemination of the accepted articles would be achieved with the help of retrained librarians, together with publishers hired for their services, who would be in charge of curating metadata in order that all articles could be properly located by search engines and downloaded for free from the Web. The governance of such service units would be similar to other research infrastructures (e.g., large telescopes, particle colliders, or supercomputers). They should be governed by three independent bodies: a scientific committee in charge of selecting the journals allowed to use the service unit for free, an executive committee in charge of designing and maintaining the infrastructure (i.e., choosing computers and hiring technical staff, such as software developers, data managers and publishing specialists), and a user committee in charge of reporting problems to be overcome and needs for better or new services.

In 2016, following the recommendations provided by the conference *Open Science - From Vision to Action* held in Amsterdam in April, Carlos Moedas, the European Com-

[15] http://openscience.ens.fr/ABOUT_OPEN_ACCESS/DECLARATIONS/2012_07_17_European_Commission_Towards_better_access_to_scientific_information.pdf.

[16] http://openscience.ens.fr/OPEN_ACCESS_MODELS/DIAMOND_OPEN_ACCESS.

[17] https://pkp.sfu.ca.

missioner for Research, Innovation and Science, announced the Amsterdam Call for Action on Open Science, stating that “After January 1st 2020, scientific publications reporting on the results from research funded by public grants provided by national and European research programs and funding bodies must be published in compliant Open Access Journals or on an Open Access Platforms”[18]. In 2017, the European Commission published the book *Europe’s Future: Open Science, Open Innovation, and Open to the World* where it is recommended that “The European Commission could then propose to declare clauses that grant exclusive rights to publishers unfair and without effect, and to force publishers to disclose these contracts. Furthermore, and consequently to Brexit, the European Commission could reconsider the present negotiation about European copyright law. Indeed, besides United Kingdom, other Commonwealth members and United States of America that are ruled by copyright, most of United Nations members are ruled by author’s law. Europe could then play a leading role to promote author’s law, to give a better protection to authors and a legal status to knowledge commons”[19].

In 2018, the European Commission, together with the association Science Europe[20] decided to accelerate the transition to full and immediate Open Access to scientific publications. This is the so-called *Plan S* which relies on the following principles:

- Authors retain copyright of their publication with no restrictions. All publications must be published under an open license, preferably the Creative Commons Attribution License CC-BY[21].
- Funding agencies will ensure jointly the establishment of robust criteria and requirements for the services that compliant high quality Open Access journals and platforms must provide.
- In case such high quality Open Access platforms or journals do not yet exist, the funding agencies will in a coordinate way provide incentives to establish these and support them when appropriate.
- Where applicable, Open Access publication fees are covered by funding agencies or universities.
- When Open Access publication fees are applied, their funding is capped (across Europe) and standardized in accordance with domain-specific requirements.
- Funding agencies will ask universities, research organizations and libraries to align their policies and strategies, notably to ensure transparency.
- Funding agencies accept the principle that all scientists are able to publish their work within Open Access even if their institutions have limited resources.
- The importance of open archives for hosting research outputs is acknowledged because of their long-term archiving function and their potential for editorial innovation.
- The hybrid model of publishing (i.e., whose subscription is paid by libraries and authors are asked to pay for their paper to be open access) is not compliant with the above requirements.

[18] http://openscience.ens.fr/ABOUT_OPEN_ACCESS/DECLARATIONS/2016_05_17_EC_Amsterdam_Call_for_Action_on_Open_Science.pdf.

[19] http://openscience.ens.fr/MARIE_FARGE/ARTICLES/2017_05_15_BOOK_CHAPTER_FOR_THE_EUROPEAN_COMMISSION/2017_05_15_Chapter_on_publishing_and_peer_reviewing_in_open_access.pdf.

[20] https://www.scienceeurope.org.

[21] https://creativecommons.org/licenses.

- Funding agencies will monitor compliance with the principles enunciated and will sanction non-compliance.

The computational biologist Michael Eisen from Berkeley University, one of the co-authors of the Budapest Declaration[22] and co-founder of PLOS, expressed his doubts concerning Plan S. He launched an Open Letter in Support of Funder Open Publishing Mandates, which has been signed by many researchers who denounce the Gold Open Access model that Plan S will impose on researchers by 2020: "We, the undersigned, are researchers who believe that the world's scholarly literature is a public resource that only achieves its full value when it is freely available to all. For too long we have tolerated a pay-for-access business model for scholarly journals that is inequitable, impedes progress in our fields, and denies the public the full benefit of our work"[23].

Today, the Green Open Access model (where authors deposit a version of their articles in an open repository) is widely developed but publishers are lobbying against it, by imposing embargos to delay deposit. Thus they make sure that only the Gold Open Access model allows immediate open access in order for their business model to win in the long-term. Only a few non commercial publishers, e.g., the American Physical Society (APS) authorizes authors to deposit the published version of their articles without any embargo period; this practice is exemplary since it avoids different versions of the same article to circulate on the Web, which is confusing because their type-setting and even their content (e.g., the author's version before peer-reviewing) are different. It already exists worldwide a large number of institutional or disciplinary open repositories, which are listed in the Directory of Open Access Repositories (DOAR)[24], where researchers can deposit a version of their articles. Researchers consider that the Green Open Access model is presently the best solution to disseminate their articles, since it ensures a smooth transition from toll access to open access, while it respects their academic freedom and leaves room for designing new publishing models, such as the Diamond Open Access model. Since embargoes reduce and distort the dissemination of peer-reviewed articles, several countries (e.g., Germany and France) have changed their copyright law to forbid embargoes or minimize their duration.

It is actually possible to overcome the publisher's embargo by providing an Open Access Button[25], which automatically sends an email to the authors and asks them to kindly provide their author version, if their article is still under embargo. Therefore the Green Open Access with an Open Access Button and the Diamond Open Access model, both designed by researchers to disseminate as widely and freely as possible their results, offer immediate Open Access and there are no longer reasons for preferring the Gold Open Access model, designed by publishers. Moreover, a way to publish in Diamond open access is to rely on the open repositories developed for Green Open Access (e.g., Zenodo[26] at CERN in Geneva), which leads to the concept of 'overlay journals', where authors deposit their article in an open repository to be peer-reviewed; for this authors have two possibilities: either they mention the overlay journal they choose for peer-reviewing their article, or they let any journal editor propose them to peer-review it. An overlay journal is simply a set of links to the articles that have been peer-reviewed and ac-

[22] https://www.budapestopenaccessinitiative.org.

[23] http://michaeleisen.org/petition.

[24] http://www.opendoar.org.

[25] https://openaccessbutton.org.

[26] http://zenodo.org.

cepted by its editorial board (e.g., *Discrete Analysis* whose articles are deposited in arXiv and then peer-reviewed by its editorial board[27]). All articles are thus in Open Access as soon as they have been deposited by their authors in an open repository and they are peer-reviewed afterwards. Moreover, any article can be copied from the open repository, which guarantees that the most useful articles will always remain available somewhere.

2.3. Open Science in France

Research in France is characterized by the fact that most research institutions and universities are publicly owned and publicly funded, in contrast to, e.g., the United Kingdom and United States. The Ministry of Research and Higher Education does not wish to let publishers control the dissemination of research results and impose the Gold Open Access model. For over thirty years the *Centre National de la Recherche Scientifique* (CNRS) was the world's largest producer of peer-reviewed articles, but in 2017 it retrograded to second rank, behind the Chinese Academy of Science. Consequently, the large number of peer-reviewed articles that CNRS researchers publish cannot be supported under the Gold Open Access model that publishers try to impose. CNRS cannot afford to pay APCs without running the risk of bankruptcy, or else should drastically reduce its production by setting a maximum number of articles to be published per year. The strategy of the major publishers is obvious: they want to impose, world-wide and as soon as possible, their Gold Open Access model, in order to preserve their profits and reinforce their control of scientific publishing. They have already succeeded in confusing most of researchers who think that Open Access implies 'author pays'; this is not true since it exists other Open Access models that scientists have designed and adapted to their needs, and that the French Ministry of Research and Higher Education is supporting.

Here are some of the main steps achieved to develop Open Science in France.

In 1999, the historian, specialist of digital humanities Marin Dacos designed the platform Revues.org and the open-source software Lodel to publish and disseminate research journals of social sciences and humanities. In 2007 he was hired by CNRS to create in Marseilles the *Centre pour L'Edition electronique Ouverte* (CLEO) and develop the publishing platform Open Edition[28], which offers standard services for free to both readers and authors (only additional premium services are charged to libraries). Up to now, Open Edition has published 6,980 books, 508 journals from 31 countries with 16 different languages, 2,926 blogs and 41,494 announcements of conferences.

In 2001, the theoretical physicist Franck Laloë, from *Ecole Normale Supérieure* (ENS) in Paris, suggested that CNRS create the open repository *Hyper Articles en Ligne* (HAL)[29] devoted to all scientific fields, including sciences and humanities, on the model of arXiv that is limited to a few exact sciences. HAL is managed by the *Centre pour la Communication Scientifique Directe* (CCSD), a service unit from CNRS located in Lyons, which ensures the long-term preservation of the deposited articles and of their metadata. HAL is moderated to check that only the author version is deposited (this verification requires a few days), and the intellectual property remains with the authors. Articles in physics, mathematics and computer sciences deposited in HAL are automatically

[27] http://discreteanalysisjournal.com.

[28] https://www.openedition.org.

[29] https://hal.archives-ouvertes.fr.

copied into arXiv, and there is a similar agreement between HAL and PubMed Central (PMC) for articles in biology.

In 2003, the director of CNRS, Bernard Larrouturou, together with the directors of several German and French public research institutions, co-signed the already mentioned Berlin Declaration, which states that: "Internet has fundamentally transformed the concrete and economic framework of the diffusion of scientific knowledge and cultural heritage [...] In the interest of our institutions, the new paradigm of Open Access must be encouraged for the benefit of science and society. [...] Our institutions must find appropriate solutions in order to let the financial and legal frameworks evolve in such a way that access and optimal use of the new facilities be guaranteed".

In 2011, Marie Farge, member of the ethics committee of CNRS, wrote the *Recommendation about relations between researchers and publishers*[30]. She explained that publishers force researchers to accept the Copyright Transfer Forms in order to give them their copyright for free (although their article passed peer-reviewing and has been accepted for publication by the journal). This is not legal under the French author's right law that differs from copyright law. Indeed, in France researchers own exclusive intangible property right on their articles that they do not share with their employer (French Intellectual Property Code, Article L. 111-1) and "The transfer of the author's rights is subject to the condition that each of the rights transferred is mentioned separately in the deed of transfer and that the field of exploitation of the rights transferred is delimited in terms of its scope and destination, place and duration.[...] The beneficiary of the assignment undertakes by this contract to seek to exploit the assigned right in accordance with the practices of the profession and to pay the author a remuneration proportional to the income received" (French Intellectual Property Code, Article L. 131-3, modified by the so-called *Loi sur le Droit d'Auteur et les Droits Voisins dans la Société de l'Information* (DADVSI) of 1 August). She suggested that CNRS, together with universities and other French public research institutions, negotiate jointly national licenses with publishers, on the model of what the Brazilian Federal State does; this recommendation has been implemented in France since 2014. She also recommended that those negotiations be conducted, not only by librarians, but also by researchers, members of editorial boards and lawyers, specialists of intellectual property law, commercial law and public market law; unfortunately, this recommendation is not yet in use.

In 2014, Antonin Delpeuch, a student in computer sciences at ENS in Paris created the platform Dissemin[31] to help researchers to deposit their articles in open repositories; his motto is: "spot your own pay-walled papers, liberate them in one click". In 2015, Antonin Delpeuch, Marie Farge and three students, all working at ENS, created the non-profit association named Committee for the Accessibility of Publications in Sciences and Humanities (CAPSH) which supports Dissemin. Presently Dissemin harvests more than 100 million scientific articles from many research fields and institutions worldwide, using various metadata sources (e.g., CrossRef and BASE). Dissemin provides researchers with a simple interface to locate and download articles already in Open Access and, for articles that are not in Open Access, it checks which version of the article (preprint, postprint or the published version) their publisher allows the author to deposit in an open repository. Presently Dissemin offers the choice between three open repositories, which

[30] http://openscience.ens.fr/MARIE_FARGE/ARTICLES/2011_06_27_2011_AVIS_POUR_LE_COMITE_D_ETHIQUE_DU_CNRS/2011_06_27_Avis_CNRS.pdf.

[31] https://dissem.in.

are well-indexed, metadata-rich and owned by a public or a non-profit institution: Zenodo funded by the European Commission[32], HAL funded by CNRS[33] and Open Science Framework (OSF) funded by the American National Science Foundation (NSF)[34]. Dissemin's code is written in Python and available for free under the open-source license Affero General Public License (AGPL), and anyone can download it from the open platform GitHub[35].

In 2017, the French policy to foster Open Science has been stated in the Jussieu Call for Open Science and Bibliodiversity[36]. Its goal is the "development of innovative scientific publishing models [...], open-source tools, [...] a secure and stable body of law across different countries to facilitate the availability of text mining, [...] national and international infrastructures which generate the preservation and circulation of contents'. It explains that its 'primary aim should be to pool local and national initiatives or to build an operational framework to fund open access publishing [...] and address the needs of the scientific community". More than 120 institutions from many countries signed it.

In 2018, the Institute of Mathematical Sciences and their Interactions (INSMI) of CNRS created the platform Mersenne[37] to peer-review and publish in Diamond Open Access research journals whose articles are formatted in LaTeX. Its guiding principles are: non-profit public service, open-source software using OJS, quality of the peer-reviewing, permanent archiving, transparency on costs and on the journal selection process. It is run by Mathdoc[38], a service unit from CNRS and the *Université de Grenoble-Alpes* (AGU) located in Grenoble. All peer-reviewing and electronic publishing services are free to readers and authors, but additional services (e.g., copy editing, proof-reading, plagiarism detection, print-on-demand) are charged. The French Minister of Research and Higher Education, Frédérique Vidal, published in 2018 the National Plan for Open Science and announced that France strongly supports the European policy proposed by the European Commissioner for Research, Science and Innovation, Carlos Moedas, requiring that by 2020 all scientific publications should be in Open Access as soon as they are published. Frédérique Vidal stated that "France is committed to ensuring that research results are open to all, researchers, companies and citizens, without hindrance, without delay, without payment". For this, she added that Open Science should be taken into account to evaluate researchers and research institutions, and she announced the creation of a special fund dedicated to support Open Science. Today, the results of publicly funded research, namely articles and data, should be by default published in Open Access since "Science is a common good that we must share as widely as possible" and "the role of public authorities is to restore the initial function of science as a factor of collective enrichment". The goal is to achieve the vision of Elinor Orstrom, a professor of political science at Indiana University, who introduced the concept of Knowledge Commons[39] and received in 2009 the Nobel prize in economic sciences for "her analysis of economic governance, especially the commons, showing how common resources can

[32] https://zenodo.org.

[33] https://hal.archives-ouvertes.fr.

[34] https://osf.io.

[35] https://github.com/dissemin.

[36] http://jussieucall.org.

[37] http://www.mathdoc.fr/centre_mersenne.

[38] http://www.mathdoc.fr.

[39] Hess, C. & Ostrom, E. (Eds.) (2007). *Understanding Knowledge As a Commons: From Theory to Practice*. MIT Press, 367 p.

be managed successfully by the people who use them, rather than by governments or private companies". It is worth recalling here that ideas are not of the same nature as material goods, because when we share an idea we do not loose it. Sharing an idea is thus a positive-sum game, and also the necessary condition for verifying and improving this idea. Hopefully, Knowledge Commons would be easier to develop than traditional commons (e.g., fisheries) since they concern ideas but not material products. Unfortunately, major publishers still control most of scientific publishing because they own the journals, created by researchers, together with the peer-reviewing reports and the articles, written by researchers. Since throughout the word research is mainly supported by public funds, given the high amount of long-term investment required, it is urgent that researchers and their funding agencies recover control of scientific publishing and develop the Knowledge Commons for preserving and sharing research output. Public funding agencies should no longer finance the APCs publishers require for Gold Access Access, but rather offer open repositories and publishing platforms for researchers to use. Indeed, together with research infrastructures necessary to produce scientific results, researchers also need publishing infrastructures to share and to preserve their publications and their data. For ensuring that commercial publishers cannot buy and control publishing platforms, as it too often happens (e.g., the platforms Mendeley, Pure and SSRN were bought by Elsevier), it is essential that those infrastructures necessary to research be publicly owned and developed using open-source software.

3. Towards Open Legal Doctrine

Is law a science like any other? Does legal research have a place in the Open Science movement? Regardless of opinions, the worlds of law and science are very close. Law is a language, an art that is coming into ever closer contact with science (AI, statistics, etc.), and is even gradually applying its methods.

One may get the feeling that the legal research community is responding to the Open Science evolution in a piecemeal fashion, sometimes even coming into apparent conflict, however progress has been constant and the needs of stakeholders in these separate worlds are seen to be converging[40].

One must also keep in mind that legal scholars have actually been practicing Open Science for a long time. The first open-access legal journals were founded in 1996, immediately following the introduction of the Internet in France, with the Neptunus university-based initiative[41] and, in the public sector, the *Cour de Cassation*'s information newsletter. Numerous other projects have followed, with over a hundred journals offering high-quality content being established over the subsequent two decades[42].

Concomitantly, as mentioned earlier, academics have been successful in updating their regulatory, then legal, environment, making it possible for authors to reference and deposit their work on specially created digital platforms similar to those created for other disciplines as part of the Open Science movement.

[40] http://bibliotheque-blogs.unice.fr/httbu/2017/01/26/gerer-et-diffuser-les-donnees-de-la-recherche-enjeux.

[41] http://www.cdmo.univ-nantes.fr/neptunus-international-884952.kjsp?RH=1342095979500&RF=1339768387194.

[42] List and short presentation of these journals on the Juriconnexion association's Wiki site https://web.archive.org/web/20170606093931/http://www.juriconnexion.fr/wiki/index.php?title=Revues_libres.

Legal literature stakeholders are faced with a real challenge, i.e. gathering together in a single digital location the enormous body of work that is published online in different forms: articles, theses, dissertations, etc. At present, there is no guarantee that all this material will be preserved. The platform that has been made available to researchers is not wholly satisfactory: depositing material could be made easier, document descriptions could be more precise, which would contribute to producing better search results; also, there is room for improvement in the official recognition and research activity ranking processes followed by authors' home institutions. By what means can law scholars be encouraged to deposit their work on a single common platform? Which would be the platform of choice? Could such a repository be improved by developing a European or international document description standard?

It is nevertheless important to note that significant progress has been achieved in the description of research work, the coordination of descriptions and their dissemination, with the creation and continuous enrichment of the Univ-Droit website[43]. The platform, which is placed under the authority of the Conference of Faculty Deans and is managed by a dynamic team, collects the work of a variety of entities: university libraries, the national cataloguing institution Agence Bibliographique de l'Enseignement Supérieur (ABES), as well as a number of publishers and university laboratories.

Mindsets in the academic world have also changed very quickly over the same period[44]. It is clear that the pathway to obtaining the agrégation en droit has hindered the free circulation of research work in the past, and still occasionally creates barriers. However, the enhanced exchange of ideas and experience and the cooperation that derive from sharing, as well as the need to reach out to new readers by disseminating work via channels other than subscription-based journals, combined with the need to see one's work referenced more frequently, have brought together the legal world and different compartments of the human science world.

Uploading articles onto public online repositories is now recommended, encouraged and supported. Thesis supervisors, librarians and administrative authorities are introducing training modules on how to deposit research work and ensure interoperability. The present revolution underway in legal training will no doubt help to consolidate this new practice. France's Pix platform[45] is already announcing new self-education tools that will facilitate submission and indexing of material. In the meantime, a number of tutorials[46] have already been made available which provide all the basic information.

This sharing of information is part of a broader movement involving other legal professionals, i.e. lawyers, notaries and bailiffs, for whom good practices in knowledge management represent a valued skill, highly sought after in the workplace. Bar associations have set up their own open access platforms and encourage their members to make their legal doctrine available to the public at large.

Publishers too are feeling the winds of change and are reflecting upon how to adapt their business models. In the past, the various players on the legal market expressed different needs, depending on whether they were undergrads, PhD students, young researchers, university professors, lawyers, or other legal professionals. Additionally, their needs would evolve as their careers would develop.

[43] https://univ-droit.fr/.

[44] https://www.village-justice.com/articles/etat-recherche-droit,28198.html?

[45] https://www.blogdumoderateur.com/pix-competences-numeriques/.

[46] https://doc.archives-ouvertes.fr/tutoriels-video/.

Today, increasingly sophisticated and powerful search engines have brought the majority of legal doctrine into the light. The Isidore search engine[47], which focuses on open-access scholarly content, is one of the most advanced tools at this time. These new, globalized tools have contributed to changing users' mindsets.

Market players more and more frequently demand having access to everything, and discrimination is less well tolerated than in the past. Publishers have to deal with an increasing demand for unified search capabilities across paid and open-access content. They have become aware that future solutions will necessarily involve joint discussions with users and authors, and will have to be future-proof.

Driven by the goal of fostering open access to knowledge, the Open Law association launched the Open Doctrine project in early 2018, as part of which a number of areas have been identified to date[48]:

- Education: inform and train all legal scholars in order to ensure that all work is deposited onto a single group of servers.
- Technology: provide a unified, reliable and lasting integration of content, and improve access.
- Institutional engagement: enlist the support of academic institutions in highlighting the value of openly accessible research content and work on leveraging content made available on open platforms.
- Economics: formalize various freemium-based data dissemination business models that blend free dissemination and pay-to-use when a service provides added value.

In order to promote this initiative and get the necessary discussion going, an Open Thesis award is to be launched in 2019[49], which will reward authors who make the full version of their legal thesis freely accessible. The award will be announced officially at a ceremony which will take place at the Cujas university library on May 15th. It is intended to encourage young researchers to share their work, and to reward actions that foster Open Science. Not only will these initiatives undoubtedly benefit young scholars' reputations, in particular in combination with social media, but they also translate a willingness to give French doctrine more international exposure, in Europe and further afield. One must, however, not lose sight of two important aspects: the language barrier and the difficulties future generations will experience in reading today's doctrine. There are numerous challenges in all areas. In law, as is the case in other disciplines, Open Science is but a new step in a long history of academic initiatives geared towards helping the research community function efficiently.

Acknowledgements

Authors acknowledge support from Open Law, ADIJ and Juriconnexion.

[47] https://isidore.science/.

[48] https://openlaw.fr/travaux/communs-numeriques/open-doctrine.

[49] https://openlaw.fr/fonds-de-dotation-open-law.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190014

Tools for Discovery: Opening Doors to Legal Research

Ginevra PERUGINELLI and Sebastiano FARO
Istituto di Teoria e Tecniche dell'Informazione Giuridica – ITTIG-CNR (Italy)

Abstract. Discovery tools are specialized portals for bibliographic research widely used in libraries with heterogeneous collections of electronic and digital resources. The Chapter provides an overview of the library resource discovery environment, explaining how these technologies, methodologies, and products might be able to adapt to changes in the evolving information landscape in scholarly communications. This Chapter also attempts to explore which are the effects of discovery tools on legal research.

Keywords. library resource discovery services, digital library, bibliographic legal research

1. Introduction

In recent years, the library world has witnessed the emergence of new bibliographic research tools, including Discovery Tools (DTs). These tools provide a single window approach to the resources subscribed by the library. This includes the Online Public Access Catalogue (OPAC), e-resources subscribed by the library, institutional repositories, open access content and many more. Similar to Google, or any other general search engine, DTs are built using a pre-harvested central index of data. Internet search engines rely on open access and public domain data to populate their central index, which can sometimes lead to broken links, inaccessible sources and dubious quality of content. By contrast, DTs use data supplied by libraries and publishers, resulting in more reliable search results with stable, direct links to licensed, full text articles and digital content [1]. Therefore, they are created to respond to users' needs of managing the new emerging class of electronic resources and of providing patrons with simpler, and web-based research services [2]; [3]; [4].

From an evolutionary point of view, they represent the effect of a long digital revolution that, starting from the 1980s, has affected the library world. Over the decades, OPACs have gone through several generational phases; it is with the birth of the Web that they acquire visibility outside the physical space of the library and become effective information retrieval tools. The Web has not only played a determinant role in the technological arena, leading a relevant number of users to prefer the use of search engines as main tools for finding information resources, but also permeating our social life [5]. The user is no longer satisfied with a traditional use of information materials, rather he/she needs to be able to move and interact with library catalogs with the same autonomy and independency used on the Web.

Gallacher [6] has highlighted a deep cultural conflict between traditionalists and the so called Google generation[1], explaining the tension between the advocates of traditional bibliographic research and those who fully rely on online research. Furthermore, the work of Gunter, Rowlands and Nicholas [7] analyzes in detail the question of a break between the pre and post Google generations. To characterize the Google generation, these authors make specific hypotheses which fully describe its bibliographic research approach: (1) prefers visual information to textual information; (2) wants a variety of learning experience; (3) has definitely adopted the digital modes of communication; (4) is multitasking; (5) is impatient and does not tolerate any delay; (6) considers its peers more credible than other points of authority; (7) needs to feel permanently connected to the Web; (8) learns more through action and knowledge; (9) prefers information in small quantities, easy to digest, rather than reading the full text; (10) has a poor understanding and lack of respect for intellectual property; (11) is not interested in format or container issues; (12) tends to put virtual reality at the same level of experience.

In the law environment, this clash is very noticeable also with respect to teaching legal research to today's law students and young lawyers. The traditionalist view of legal research is essentially based on the belief that law research based on traditional bibliographic sources such as books, commentaries, print journals is superior to online research, at least as a first step in research. On the other hand, it has been observed that more and more current legal generation is so online-oriented that it has learned to rely on powerful search engines to provide answers to complex questions. Young jurists refer more and more to search engines to conduct complex and comprehensive search pointing to different resources (not just bibliographic) with one stop search. The easy-to-use appeal of the Web for legal research is attracting more and more followers fascinated by the idea of finding the solution to a practical legal case. Furthermore, most students entering universities are more comfortable working on a keyboard and reading from a computer screen rather than using paper in their hand.

In this context and in the other sciences as well, DTs carry out their task reconciling different needs and becoming the undisputed leader in the field of bibliographic research [8]. On the one hand, DTs are closely related to the evolution of libraries and to the quality services they offer; on the other, they are meeting habits of the Google generation.

This Chapter provides an overview of the actual resource discovery environment with a particular focus on the advantages and disadvantages of these tools in the library world. Authors conclude with some reflections on possible developments and effects of these tools on legal bibliographic research.

2. Origin of Discovery Tools

In recent years two innovative types of bibliographic research tools have been developed: the Next Generation Catalog (NGC) and the DT. They represent an evolution with respect

[1] The 'Google generation' is an expression referring to a generation of people, born after 1993, growing up in a world dominated by the Internet. The expression has entered popular usage as a shorthand way referring to a generation whose first port of call for knowledge is the Internet and a search engine, Google being the most popular. This is used in contrast to earlier generations who gained their knowledge through books and conventional libraries.

to OPACs, which since the beginning of the catalog automation era have been the main bibliographic research tools for information retrieval of library collections.

OPACs have been evolved over time: from simple information retrieval tools allowing only a few types of research, to third-generation OPACs, enhanced in functionality and usability, expanded in indexing, data records and collection coverage and extended through links and networks, acting as a gateway to additional collections.

After the mid-2000s, a strong dissatisfaction with OPACs began to appear. In that period the Web evolved at technological and functional levels. The user becomes accustomed to new search engines, also utilized as information retrieval tools for bibliographic research.

A first response was the NGC, which has been defined as an application that still uses data in traditional form, allowing users to perform a simplified search based on friendly interfaces similar to the Web while integrating external resources. The use of these tools spread between 2005 and 2007. The aim was to achieve greater integration within the Web, trying to provide ease of use, better ways to communicate, search and use information. The new features and services introduced by NGC concerned data, search, information retrieval and displaying of results.

Starting from 2007 the NGC begun to lose their primacy in favor of DTs for bibliographic research. In 2010, Marshall Breeding announced that NGCs would become unsuitable and obsolete instruments and, for this reason, they would have been overthrown by DTs [9]. At their origin, there is an increase in digital publishing and the consequent need to manage the huge amount of electronic resources that lead libraries to adopt a tool capable of adequately representing the new type of electronic resources such as e-books, e-journals and databases, while offering a simultaneous access to different types of bibliographic resources.

3. Structure, Components and Functionalities

The structure of DTs is organized into three parts [8]: (i) index, (ii) search interface, and (iii) link resolver.

The *index* includes metadata and full text resources resulting from agreements with commercial publishers, to which are added metadata and open access resources also contained in research repositories. The central index is the main element of competition among the producers of this kind of tools [10]. The amount of indexed metadata depends on existing agreements and licenses between libraries and digital content providers. It is worth mentioning that access rights to metadata and full texts are governed by two types of agreements, independent of each other: those stipulated by the institution and those stipulated by the DT producer adopted by the library.

The *search interface* is similar to search engines' user interface. Navigation is allowed without the need for special instructions, starting from a unique Google like string, or with the possibility of articulating multiple search keys in an advanced search modality; the results are then presented in a short or complete form. Ranking of results by relevance is a common feature of the various DTs and is combined with other sorting criteria, which, however, vary according to the product chosen by the library. The algorithms that allow sorting by relevance of search results are owned by the DTs' vendor and are not always made public [11].

Link resolvers are software that use OpenURL standard[2] and a knowledge base to connect a searcher from a citation to the item held in the library or to online full text content. If the content has been subscribed by the library, the link resolver will directly connect the searcher to the subscription content site. Link resolver software and their associated knowledge bases are essential technologies for modern academic libraries. The success of link resolvers is dependent on (i) complete, consistent, accurate citation metadata, (ii) well-defined knowledge base holdings, and (iii) accurate link syntax as generated by the software. Selecting link resolver and knowledge base software has become more complicated than ever, due to the increasing need for system interoperability. Libraries' expectations for clean metadata following professional and industry standards have correspondingly increased [13].

As a matter of fact DTs are index-based systems: the content of databases, local and remote, is re-indexed by these tools. During the indexing process, the system should treat all content equally. It is not clear how content is indexed and, therefore, recovered in the research phase: there are no standards regulating this process [8] and problems possibly arising from reindexing derive from the different quality and completeness level of metadata that these discovery systems receive from different sources, where metadata can already vary in terms of quality and quantity [14]; [15].

On the basis of this structure, the core features of these tools are represented by the content and technology used. Content, as knowledge base, includes: journals in any format, books, databases, aggregators' content, open source materials, newspapers, local indexes (library catalogue, institutional repositories). Technology includes, among others: harvester (OAI/PMH etc.), automated transfer routines, metadata mapping, indexing technology, de-duplication algorithms, link resolvers, relevancy algorithms, interface technologies.

These components allow DTs to provide users with a great array of functionalities [16] which are briefly described in Table 1. Of course the functionalities showed are included in the majority of most popular library discovery services software available on the market today. A list of selected proprietary and open source software is shown in Table 2.

These are only some of the functionalities that libraries look for. Other important components include ease of implementation, price, vendor support and estimated content coverage.

It is to be noted that the availability of open source solution affects the price charged by commercial vendors. At the same time, the implementation cost of open sources software must be taken into consideration as it requires a lot of expertise on the part of libraries which want to implement it or decide to depend on external service providers for implementing it.

4. Pros and Cons of Discovery Tools

Ease of use seems to be one of the main advantages of DTs and appears as the answer to Breeding's call for a seamless experience presenting a consistent interface, despite the

[2] As stated in [12] OpenURL framework provides "a standardized format for transporting bibliographic metadata about objects between information services". NISO standardized the OpenURL protocol in 2004 as ANSI/NISO Z39.88 and many vendors developed and released their own link resolver software.

Table 1. Advanced features library looks for in DTs

Functionality	Description
One stop search	All library resources in one search
Modern design web interface	A design similar to general search engines
Enriched content	Book, images, reviews, user driven input, such as comments, descriptions, ratings, and tag clouds covered
Faceted navigation	Users are allowed to narrow down search results by categories, also called facets (location, publication date, author, format...)
Simple keyword search box	A simple keyword search box of a Google like type is offered. An advanced search is also provided
Results ordered by relevance	Relevance algorithms are applied to the list of results
Recommendations/related materials	Suggestions like 'readers who borrowed this item also borrowed the following ... or a link to recommended readings are offered
Integration with social networking sites	Users can share links to library items with their friends on social networks
Persistent links	A stable URL capable of serving as a permanent link to the record is available

Table 2. Selection of proprietary and open source software

Proprietary tools	EBSCO Discovery Service (EDS); Summon (ProQuest); Axiell Arena 3.1. (Axiell); BiblioCore (Biblio Commons); Primo and Primo Central Index (ex Libris); WorldCat Local (OCLC); OvidDiscovery; Inspire™ Discovery (Innovative Interfaces); Endeca (Oracle); Enterprise (Sirsi Dynix)
Open source tools	Blacklight (University of Virginia Library); VuFind (Villanova University); eXtensible Catalog/XC (University of Rochester); LibraryFind; Franklin; SOPAC

use of multiple technology and content products behind the scenes [17]. According to Gross and Sheridan's studies and usability tests [18], students certainly find DTs an easy way to get results, probably easier than the various options they were confronted before.

In addition, users will be searching a much larger data set than previous databases were able to offer. As Vaughn points out, these new services, with hundreds of millions of items, many of them full text, previously housed in dozens or hundreds of individual silos [19], make it easier for users to find content that they would otherwise miss.

Furthermore, as stated by Way [20] in its usage statistics, after the implementation of a DT, the sharp drop in single database usage, associated with high increase in the number of full-text downloads and resolver clicks of links, suggest that such tool has a considerable impact on users' search behavior and on use of library collections.

However, there are also some relevant disadvantages [21] which are very considerable when libraries choose to rely on DTs.

The number of results and varied formats that DTs return to users is overwhelming, especially on simple, non specific searches [22]. The huge number of results coupled with the increasing amount of available object types and formats can make for a confusing jumble of results. One of the fields in which DTs find greater use are university libraries where there is a greater need to offer a unique search point that coordinates access to materials such as electronic resources and databases through a single tool, capable of managing the authentication to resources reserved for the users of an academic institution

[23]. In such a context, most users need assistance to refine their search. This massive number of records is also a concern for specialty librarians. In many cases, users are interested in a specific set of resources, and the use of DTs does not fit their information needs. Moreover, from an empirical research it has emerged that the students "once they had chosen [the web scale discovery tool] ... stayed with it even when an alternative pathway may have proved more fruitful" [18].

Cost issues are also a concern, as DTs, both from a monetary and staff time perspective, require lots of resources. Initial design and setup of the system, testing, and implementation require time and a specialized skill set.

Another disadvantage regards unrealistic user expectations. Several librarians expressed concern about the possible creation in students' mind that these tools set "the expectation that everything is available online in full text". It was also found that some users expressed frustration and disappointment when the tool pointed them to a physical book, located in the stacks at their library [24].

Finally, in order for content to appear in DTs, it must be licensed by both the library and the DT vendor. This leads to situations where only portions of a library's holdings are searchable via DTs. Users may still need to be directed to conduct their search in topic-specific databases. For this reason DTs require a coordination between content producers, resource discovery service providers and libraries.

Therefore, it is up to libraries, in collaboration with their partners, to set up the central index (which resources will be indexed, etc.), the link resolvers to make the best use of the resources available to their users, and integrate the authentication systems in a transparent manner. It follows that the value of the service will largely depend on the coordination work between content producers, software solution providers, and libraries.

5. Possible Developments

To reflect on what is missing and on the opportunities that DTs can take, the work of Marshal Breeding, helps us [8]. We resume, from his contribution, some of the features not fully realized in the current generation of DTs.

DTs well represent relevant material, but omissions in coverage remains. As an example, resources redacted in non-English languages are not covered in an optimal way. In particular, in the field of law, some of these developments would be very useful for improving legal bibliographic research. Articles, books, journals in many areas of law refer to specific national legal orders and have been published in the language of the jurisdiction analyzed by the authors. Multilingual search represents a crucial issue for the future development of DTs so that the content represented in the discovery index becomes more and more heterogeneous by language.

Advanced and precision searching remains a very important functionality for DTs and a consistent management of metadata should be a major step towards enhancement of these components.

Also the exposition of open access materials from a variety of sources is a big challenge. These resources are very numerous and DTs often provide duplicate and unclear answers. In fact, the original metadata of open access resources often do not follow any standard: DTs do not worry about deduplication [10].

Furthermore, DTs should strongly consider legal blogs and social media, able to quickly summarize the development and evolution of legal science. Many journal articles mention as bibliographic references blogs of scholars and professionals. The community of legal professionals is getting closer and closer to these containers that are quickly spread for their simplicity of implementation and immediacy of communication. Some bloggers with time have become opinion makers and influencers, constituting an alternative source to traditional legal ones, with which they sometimes collaborate, while maintaining the dialogic character with readers.

Relevancy ranking is another key issue to consider for the developers of DTs. However, how relevancy works is ignored by librarians and users.

Enhanced discoverability through non-textual associations is also desirable. Clustering technologies may be able to produce facets based on the content of resources retrieved to guide the user towards the ones that match his/her interests. Progress has been made, but there is still much room for improvement.

Nevertheless, Breeding continues by listing future enhancements that should be made in response to requests from libraries and users. In particular, some of these improvements are related to social features, analytics, altmetrics. Opportunities to enable social interaction would depend on standardized mechanisms that enable interoperability between the ecosystems of discovery services and those of external social networks. Furthermore, libraries and publishers have considerable interest in the ability to measure the performance of their discovery services and which resources have been retrieved. Finally, alternative measures relating to the description of the impact of scholarly resources and the performance of academic libraries are undoubtedly necessary. It can be discussed to what extent they can become part of the discovery ecosystem and whether they can be used in relevance algorithms to help identify materials of higher interest or quality.

6. The Impact of Discovery Tools on Bibliographic Legal Research

DTs have determined a paradigm shift. They give greater importance to discovery than to simple search typical of information retrieval techniques on which previous tools were based [25].

DTs offer a different approach to information, aiming at providing awareness, rather than a specific response, by aggregating content related to a particular area of interest.

In such a way, the path of discovery follows a more explorative approach, often driven by a generic need, or at least less explicitly stated. Actually, DTs are born with the goal of providing the certainty of not missing the most important information on a specific topic.

In the legal domain, this approach is not always the correct one. As a matter of fact, legal users are of different types and often DTs are not appropriate for all these kind of users. If we think of a scholar expert in a specific field and legal concepts, he/she has a different approach of searching that does not necessarily require the support of a DT. More advanced users may find discipline-specific databases still useful in providing a better search performance. Instead, a student approaching a general topic for the first time, needs DTs to get a rough outline of the topic. However, centralized index-based DTs can obscure the complexity of the information retrieval process because of the immediacy in retrieving the resources proposed. Immediate access to full text may prompt to download a result in full-text just because it is available [26]; [27]. This also applies if we move on the types of libraries. A university library, if equipped with a DT will offer to its users a useful service for legal research; whereas in specialized libraries which cover specific areas of law, the library catalog and the resources selected by librarians (databases, ejournals ...) could fully satisfy a legal expert.

Therefore, the exploitation of DTs is certainly only one phase of the legal research process, which is a complex task that is not resolved in the exploration phase of the resources, but requires the evaluation, interpretation and connection with other sources relevant to the study of law. The real solutions to complex legal issues and problems require analysis and thoughtful conclusions. Each research question has a different starting point, process, and conclusion. Of course, it includes false starts, dead ends, and revisions. Most important, legal research is never 'stopped' but the skilled researcher recognizes when to finish. As Felix Frankfurter stated [28]: "Research requires the poetic quality of the imagination that sees significance and relation where others are indifferent or find unrelatedness; the synthetic quality of fusing items theretofore in isolation; above all the prophetic quality of piercing the future, by knowing what questions to put and what directions to give to inquiry".

A successful researcher is one who understands how to use the many resources that are available in a flexible and efficient manner.

One last point concerns the difficulty to predict the future of DTs as libraries continue to struggle to find their path in the actual shifting environment of information provision. However librarians are the main actors for undertaking action and have a fundamental role to play in selecting, implementing, and evaluating the appropriate DTs for specific contexts, as well as in training users to exploit these tools effectively, helping them to interpret the results obtained. Linking these tools to the library computing environment is also a crucial point in their implementation and a prerequisite for their proper functioning. For sure, in the law domain information professionals are uniquely placed to shape and lead all necessary changes needed in bibliographic legal discovery.

References

[1] Wang, Y. & Mi, J. (2012). Searchability and Discoverability of Library Resources: Federated Search and Beyond. *College & Undergraduate Libraries, 19*(2-4), 229-245.

[2] Morgan, E. L. (2006). A "Next-generation" Library Catalog - Executive Summary. *LITA Blog: Library Information Technology Association*, 7 July 2006, http://litablog.org/2006/07/a-next-generation-library-catalog-executive-summary-part-1-of-5/.

[3] Nagy, A. (2011). Defining the Next-generation Catalog. *Library Technology Reports, 47*(7), 11-15.

[4] Yu, H. & Young, M. (2004). The Impact of Web Search Engines on Subject Searching in OPAC. *Information Technology and Libraries, 23*(4), 168-180.

[5] Coyle, K. (2007). The Future of Library Systems, Seen from the Past. *Journal of Academic Librarianship, 33*(1), 138-140.

[6] Gallacher, I. (2006). Forty-Two: The Hitchiker's Guide to Teaching Legal Research to the Google Generation. *Akron Law Review, 39*(1), https://ideaexchange.uakron.edu/akronlawreview/vol39/iss1/5.

[7] Gunter, B., Rowlands, I. & Nicholas, D. (2009). *The Google Generation. Are ICT Innovations Changing Information-seeking Behaviour?* Chandos Publishing, 299 ff.

[8] Breeding, M. (2015). *The Future of Library Resource Discovery*. NISO, http://groups.niso.org/apps/group_public/download.php/14487/future_library_resource_discovery.pdf.

[9] Breeding, M. (2010). *Next-gen Library Catalogs*. Neal-Schuman.

[10] Trombone, A. (2019). Formare e gestire collezioni con i discovery tools. *Biblioteche oggi, 39*, 13 and 18.

[11] Breeding, M. (2013). Next-generation Discovery: An Overview of the European Scene. In Chambers, S. (Ed.). *Catalogue 2.0. The Future of Library Catalogue*. Facet Publishing, 37-64.

[12] Van de Sompel, H. & Beit-Arie, O. (2001). Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib Magazine, 7*(3).

[13] Chisare, C., Fagan, J. C., Gaines, D. & Trocchia, M. (2017). Selecting Link Resolver and Knowledge Base Software: Implications of Interoperability. *Journal of Electronic Resources Librarianship, 29*(2), 93-106.

[14] Somerville, M. M. (2013). Digital Age Discoverability: A Collaborative Organizational Approach. *Serials Review, 39*(4), 234-239.

[15] Biagetti, M. T., Schaerf, M., Iacono, A. & Trombone, A. (2017). Verifica della disponibilità delle monografie attraverso i cataloghi delle biblioteche (VerDiMAC). *ANVUR Working Papers No. 4*, http://www.anvur.it/wp-content/uploads/2017/06/WPS%20201704-VerDiMAC.pdf.

[16] Yang, S. Q. & Li, L. (2015). *Emerging Technologies for Librarians: A Practical Approach to Innovation*. Chandos Publishing, 1st ed., 31-32.

[17] Breeding, M. (2010). The State of the Art in Library Discovery 2010. *Computers in Libraries, 30*(1), 31-35, http://www.librarytechnology.org/ltg-displaytext.pl?RC=14574.

[18] Gross, J. & Sheridan, L. (2011). Web Scale Discovery: The User Experience. *New Library World, 112*(5-6), 236-247, https://doi.org/10.1108/03074801111136275.

[19] Vaughan, J. (2008). Investigations into Library Web-Scale Discovery Services. *Information Technology and Libraries, 31*(1), 32-82, https://doi.org/10.6017/ital.v31i1.1916.

[20] Way, D. (2010). The Impact of Web-scale Discovery on the Use of a Library Collection. *Serials Review, 36*(4), 214-220, http://dx.doi.org/10.1016/j.serrev.2010.07.002.

[21] Hoy, M. B. (2012). An Introduction to Web Scale Discovery Systems. *Medical Reference Services Quarterly, 31*(3), 323-329.

[22] Machetti, C. (2016). Biblioteche e discovery tool: il caso OneSearch e l'ateneo di Siena. *AIB studi, 56*(3), 395.

[23] Johnson, M. (2013). Usability Test Results for Encore in an Academic Library. *Information Technology and Libraries, 32*(3), 59-85, http://ejournals.bc.edu/ojs/index.php/ital/article/view/4635.

[24] Howard, D. & Wiebrands, C. (2011). *Culture Shock: Librarians' Response to Web Scale Search*, http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=7208&context=ecuworks, 9.

[25] Marchitelli, A. (2014). Il catalogo connesso. *Biblioteche oggi, 32*(6), 13.

[26] Breitbach, W. (2012). Web-Scale Discovery: A Library of Babel? In Pagliero Popp, M. & Dallis, D. (Eds.). *Planning and Implementing Resource Discovery Tools in Academic Libraries*. IGI Global, 637-645.

[27] Guthrie, A. & Mccoy, R. (2014). A Glimpse at Discovery Tools within the HBCU Library Landscape. In Spencer, J. S. & Millson-Martula, C. (Eds.). *Discovery Tools: The Next Generation of Library Research*. Routledge/Taylor & Francis, 177-191.

[28] Frankfurter, F. (1930). The Conditions for, and the Aims and Methods of Legal Research. *Iowa Law Review, 15*, 124.

This page intentionally left blank

Section II.2

Interoperability and Standards

This page intentionally left blank

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190016

The European Legislation Identifier

Thomas FRANCART, John DANN, Roberto PAPPALARDO
Carmen MALAGON and Marco PELLEGRINO [1]
Publications Office of the European Union

Abstract. The European Legislation Identifier initiative (ELI) aims at bringing legislation into the global Web of data, to facilitate the access, sharing and interconnection of legal information. It proposes the creation of URI identifiers for legislation based on common components and the description of their metadata based on an ontology relying on FRBRoo; the ELI ontology includes in particular the description of the FRBR levels of abstraction, the needed date properties to describe legislation and links to relate legislative acts. Legislation metadata is thus viewed as a global graph of interconnected entities. While ELI tries to lower the entry barrier for legal publishers to disseminate structured metadata and currently counts 13 implementations, it is also facing challenges to progress towards its full potential: data quality, description of ELI datasets, alignment of thematic vocabularies or granular description of the text subdivisions. ELI has the potential to facilitate access to legal information by enabling unambiguous legal citations mark-up, giving legislation more visibility in major web search engines, describing early legislation drafts or facilitating the exchange of data between legal information systems. ELI is tightly connected to novel legal information system architectures, based on legal knowledge graphs; this style of architecture encourages legal publishers to move from a document-centric perspective towards a data-centric perspective, as exemplified by the Casemates in Luxembourg and the Cellar at the Office of Publications of the European Union.

Keywords. ELI - European Legislation Identifier, web of data, ontology, FRBRoo, knowledge graphs, interoperability

1. The Web of Data, Legislation, and ELI

1.1. A Web of Data, but Where Is the Web of Legal Data?

The Web of data, or semantic Web, is a set of standards and principles that defines a new paradigm to make structured data interoperable on the World Wide Web. These principles are successfully applied to share, amongst others, libraries or open data portal catalogues, structured data for web search engines, medical databases, geographic atlases, and more.

However, legislation is not yet fully part of this Web of data, hence corresponding machine-readable descriptions cannot be reused or linked to, nor are they interoperable at web scale.

The conclusions of the Council of the European Union inviting the introduction of ELI state that "*... a European area of freedom, security and justice in which judicial cooperation can take place requires not only knowledge of European law, but also mutual*

[1] Author: T. Francart; co-authors: J. Dann, R. Pappalardo; contributors: C. Malagon, M. Pellegrino.

knowledge of the legal systems of other Member States, including national legislation"[2]. The exchange of legal information is key in this regard, and it is time that legislation becomes integrated into the Web of data, for higher interoperability.

1.2. ELI: Better Accessibility of Legislation Information through the Web of Data

From this rather technical perspective, the European Legislation Identifier initiative (ELI) aims at bringing legislation into this ocean of interconnected data. This means, from an end-user perspective, that ELI will *facilitate the access, sharing and interconnection of legal information* published through national, European and global legal information systems. This is done with the following benefits in mind:

- *Easier access* to legislation for end-users.
- Development of *new services* through the smart reuse of data.
- *Cost savings* for publishers.
- Higher *quality and reliability* of data, based on review and feedback from data reusers.
- Increased *transparency* for citizens and watchdog organisations.
- Improved *interoperability* of legislative information across legal information systems.

1.3. ELI: Motivations and Current Status

First established in the context of the European Forum of Official Gazettes[3] and with the impulse of Luxembourg, ELI has been further supported by the subgroup mandated by the Council of the European Union in the framework of the Working Party on E-law[4]. The first Council Conclusions on the European Legislation Identifier were published in 2012, and revised in 2017. The initial motivations were the ability to reuse automatically the structured description of EU Directives in national systems, as well as the need to have a shared interoperable model to link legislation on the Web; hence the use of semantic Web technologies was foreseen from the beginning.

As of March 2019, ELI has been implemented by Austria, Belgium, Denmark, Finland, France, Ireland, Italy, Luxembourg, Portugal, Spain, Norway, the United Kingdom and the EU Publications Office. The details of each implementation are available in the ELI registry[5].

2. The ELI Framework

2.1. The 3 Pillars of ELI

To have legislation dive into the data-driven world, the formal specifications[6] of ELI advocates the following:

[2] OJ C 441, 22.12.2017, p. 8-12, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017XG1222(02), with the initial version of 2012 to be found in OJ C 325, 26.10.2012, p. 3-11, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XG1026(01).

[3] https://publications.europa.eu/en/web/forum_official_gazettes/.

[4] http://www.consilium.europa.eu/en/council-eu/preparatory-bodies/working-party-e-law.

[5] https://eur-lex.europa.eu/eli.

[6] OJ C 441, 22.12.2017, p. 8-12, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017XG1222(02), with the initial version of 2012 to be found in OJ C 325, 26.10.2012, p. 3-11, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XG1026(01).

- Give stable web identifiers to legislation, using URIs; The URIs are formally described by templates, using semantic components from a legal and an end-user point of view, making them as close as possible to how users cite legislation and therefore user-readable. The use of web identifiers enables web-wide linking of legislation.
- Describe legislation metadata in a standardised way, using a common ontology.
- Make legislation metadata available for machines on the Web, by embedding structured data in web pages using RDFa[7] *or JSON-LD*[8].

The following Sections will provide a brief outline of the ELI components, focusing on what makes the added value of ELI.

2.2. *ELI Identifiers*

The ELI identifiers are made of formal components that ELI publishers can arrange in any order to specify their own URI patterns[9]. ELIs are crafted to be stable over time and serve as permalinks to legislation, to be transparent for a human reader and to be associated with a user behavior when a user links to it. As an example, http://data.europa.eu/eli/dir/1980/181 is the identifier of EU Directive 80/181/CEE, and returns the latest consolidated version of that directive.

2.3. *ELI Ontology*

While interested readers can refer to the detailed documentation of the ELI ontology[10] and the resources to implement ELI[11] for a detailed understanding of the ELI ontology, we describe 3 key features of the ontology here: its FRBR structure, specific legal dates and relations between entities. We emphasize that the ELI ontology is under constant improvement since its inception, currently in version 1.2, with a version 1.3 foreseen in late 2019.

2.3.1. *FRBR: From the Resource Paradigm to the Graph Paradigm*

The ELI ontology defines how legislation metadata must be structured in the context of the ELI framework. The ontology is based on the paradigm introduced by the Functional Requirements for Bibliographic Records conceptual model (FRBR)[12], and more specifically with its object-oriented derivative FRBRoo[13]. While legacy approaches, in bibliographic descriptions like Dublin Core Terms[14], were centered around the description of a ‘flat’ resource, FRBR splits the resource into layers of abstraction, organised hierarchically. Each layer is described with specific metadata, can refer to other entities and can be linked to, thus turning legislation information into a graph (it should be noted however that ELI has always retained a compatibility with Dublin Core by explicitly mapping its

[7] RDFa: https://www.w3.org/TR/rdfa-core/.

[8] JSON-LD: https://json-ld.org/.

[9] Reference ELI template components: http://publications.europa.eu/resource/cellar/c2f0e4f9-ed6f-11e8-b690-01aa75ed71a1.0001.03/DOC_2.

[10] https://publications.europa.eu/en/web/eu-vocabularies/eli.

[11] https://eur-lex.europa.eu/eli-register/resources.html.

[12] FRBR: https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

[13] FRBRoo: http://www.cidoc-crm.org/frbroo/.

[14] Dublin Core Terms: http://dublincore.org/documents/dcmi-terms/.

metadata fields to the corresponding Dublin Core properties, for informative purposes only).

The same paradigm influences other initiatives in the structuring of legal resources, such as Akoma Ntoso[15] or CEN Metalex [1].

In line with FRBRoo, ELI has specified the following levels of abstraction, from the more abstract to the more tangible:

- The Legal Resource covers 2 notions: a uniquely identified piece of legislation, independent of its version, language or file format; e.g. Directive 80/181/CEE (an FRBRoo Complex Work). And also a specific temporal version of a piece of legislation; e.g. version consolidated on the 27/05/2009 (a FRBRoo Individual Work).
- The Legal Expression: a linguistic variant of a version of the item of legislation; e.g. Hungarian translation.
- The Format (equivalent to FRBRoo Manifestation Product Type): a given file or set of files in a specific format, containing the written encoding of (a given version in a given language of) a piece of legislation; e.g. the body and annexes PDFs of the Hungarian translation of the consolidated version of the directive.

This graph data structure makes it possible to refer precisely to a given level, depending on the context: ‘Article 1 of Regulation No 561/2006’ (Abstract Legal Resource), ‘Regulation (EC) No 561/2006 of the European Parliament and of the Council of 15 March 2006’ (Legal Resource), ‘Students will study the German translation of Regulation No 561/2006’ (Legal Expression), ‘I have downloaded the HTML file of German translation of Regulation No 561/2006’ (Format). In particular, in a pure resource/document centric system, a reference to the ‘Abstract Legal Resource’ level is not possible.

2.3.2. Dates of a LegalResource

ELI also specifies, among many more metadata, the dates needed to describe a LegalResource ‘lifecycle’:

- Date of document: the date on which the text became a law, e.g. by the virtue of a signature by the head of state; this is different from the date on which the text was written.
- Date of publication: the date on which the text was published in an Official Journal (OJ), which may typically happen a few days after the text officially became law.
- Date range in which the legislation is in force: the time span during which the legislation is in force.
- Date of applicability: the date on which the legislation becomes applicable; this can differ from the date the legislation becomes in force (e.g. if an act in force states it will become applicable in 3 months).

2.4. Links to Create a Graph

The ELI ontology provides links to specify how legal entities can be connected together in a graph. A legal resource may simply *cite* another one. Some of these links are related to how legal provisions affect other pieces of legislation: a legal resource may *amend*, *repeal* or *commence* another one. A legal resource can also *correct* another legal resource,

[15] Vitali, F. (2007). *Akoma Ntoso Release Notes*, http://www.akomantoso.org.

like corrigenda in EU legislation, when it is only correcting spelling mistakes with no impact on the legal content. Secondary legislation is *based on* primary legislation. In a European context, a national legal resource may formally *transpose* an EU Directive. A legal resource may be *another publication* of the same resource already published elsewhere (typically in the context of national and regional OJs). And finally a legal resource may be *cited by* a case law.

2.5. The ELI Philosophy

As a complement to the description of the ELI framework, we think it is important to outline some of the key ideas that motivated its design.

Probably, the most fundamental of these ideas is to *lower the entry barrier for ELI implementations* as much as possible. Official Journals with limited resources and without in-depth technical knowledge of semantic Web principles should be able to implement ELI. The goal is to reach a critical mass of adoption and available data, so that ELI becomes a cornerstone for legal data interoperability. This is what guided the choice of RDFa as a dissemination technique for metadata, which is far less complicated than opening a SPARQL service.

In addition, ELI is *non-intrusive* with respect to existing legal publishing systems. Since ELI is a framework for the dissemination of metadata, it does not require any change of the publishing workflow.

However non-intrusive, the *progressive* nature of ELI is also an opportunity for an official legal publisher to improve the way its information is published. This can be in terms of quality, quantity or structure of information maintained, or in terms of the publishing system as a whole, by adopting a legal knowledge graph architecture (see below).

Another key aspect of ELI is its *adaptability to different legal systems*; in particular common law systems and civil law systems. This adaptability has a concrete consequence on the ontology: publishers may consider subsequent versions of a legislation either as FRBR LegalExpressions of the same LegalResource, or as independent LegalResources grouped under the same Abstract LegalResource. This last choice is adopted by most ELI implementers.

3. ELI Challenges

What are the challenges that ELI is currently facing to progress towards its full potential?

3.1. Data Quantity and Quality

Currently, 12 national legislation publishers, plus the Publications Office of the EU, have implemented ELI. However, not all implementations have the same level of precision; this is the other side of the coin for the low entry barrier principle described above: some populate a lot of metadata fields in their structured descriptions of legislation, while others provide very few, or sometimes only a subset of their legislation. This currently makes it hard to implement a ‘cross-national’ search on European legislation based on ELI metadata. Opening publicly these structured metadata may have a positive impact on their overall quality, if a feedback loop from data reusers is established.

3.2. Aggregation of ELI Datasets

While the inclusion of metadata inside webpages is the simplest path for a data provider with an existing web portal to disseminate reusable data on the Web, this does not make the life of data consumers easier. They need to crawl and fetch every webpage, analyse their HTML source in order to recreate a complete graph that can be queried and integrated into an application. While this is technically feasible, such a mechanism raises the entry barrier for data consumers and, more problematic, it may not give the guarantee to obtain the complete set of metadata from a given ELI publisher, depending on the crawling algorithm used. This is why ELI has proposed a methodology for ELI providers to describe and disseminate their ELI dataset[16], in order to facilitate the acquisition of ELI metadata by reusers. ELI should address this relation with data reusers in order for the initiative to be entirely fruitful.

3.3. Thematic Vocabularies Alignment

A key entry point for accessing legislation is the search on thematic/subject keywords. The ability to perform such a cross-national search requires first of all that legislation is indexed on a controlled set of concepts, and secondly that these sets of concepts are aligned with a pivot vocabulary allowing a higher degree of interoperability of the legal notions used to index legislation. A first step in this direction has been conducted to align the thematic vocabulary of Luxembourg with Eurovoc[17], the multilingual and multidisciplinary thesaurus covering the activities of the EU, based on either lexical proximity or on the analysis of how directives and their national transpositions are respectively indexed[18].

3.4. Granularity (Identification and Description of Subdivisions)

Linking a piece of national legislation with the EU Directive it transposes is good, but linking *articles* of that legislation with *articles* of the directive being transposed would be even better. This requires an identification of each text fragment, at EU and national level. This will be particularly useful for long directives (e.g. Directive 2009/138/CE[19] has 312 articles in 155 pages). While there is no foreseen difficulty in modelling the subdivision metadata, the challenge resides in adding complexity and scalability requirements to existing dissemination systems.

3.5. Disseminating ELI for Regulation Agencies

Use-cases have pointed out that ELI would be even more useful if it could be applied to the texts of regulation agencies [2]. The first application of ELI on *soft law* is the implementation at the French 'Autorité des Marchés Financiers'[20] (AMF): this has proven that ELI is applicable outside the scope of Official Journals. How can such agencies engage even more with ELI in order to produce machine-readable metadata and link it to legislation?

[16] https://eur-lex.europa.eu/content/eli-register/ELI_dataset_description-EN.pdf.

[17] Eurovoc, maintained in the EU vocabularies portal at http://eurovoc.europa.eu/.

[18] Dann, J. & Gerencsér, A. (2018). *EU Vocabularies – Facilitating the Linking of Legal Data*. Slides presentation at LVI 2018 Conference, https://bit.ly/2W3dFfx.

[19] http://data.europa.eu/eli/dir/2009/138/oj.

[20] https://www.amf-france.org/eli/fr/aai/amf/rg/20190209.

4. ELI Potentialities

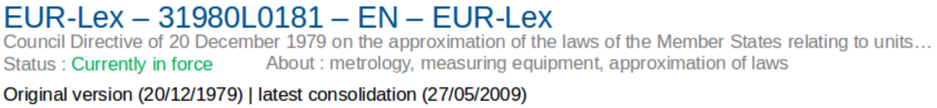
ELI has the potential to change the legal world and facilitate access to legal information, not only for end-users but also between legal information systems.

4.1. Unambiguous Legal Citations Mark-up

As an ELI identifier is crafted to be as close as possible to how an end-user cites legislation, it provides an easy way to associate textual citations with the corresponding stable web identifier to navigate to this text. ‘Article 2 of Directive 2009/138/CE’ can be easily (and automatically) converted to http://data.europa.eu/eli/dir/2009/138/art_2 (the conversion needs to know that ELI identifiers for EU legislation all start with http://data.europa.eu/eli). This allows the automated mark-up of legal citations with a higher degree of reliability, as recently shown in an effort to generate legal mark-up – including citation mark-up – on 5 Luxemburgish codes[21].

4.2. Better Visibility of Legislation in Major Search Engines

Facilitating access to legislation for end-users is one of the goals of ELI. This includes, of course, accessing legislation from the Web in general. But how do users search for information on the Web: do they use the search bar of their browser, the search field of major search engines, or the chatbot of their own operating system? These major players consume structured data from the Web to improve their services, however they do not consume just *any* structured data, the data has to be expressed in a specific and commonly agreed vocabulary: schema.org[22]. In order to make structured legislation data available for these major players, the ELI Taskforce[23] proposed to introduce a description for legislation inside this vocabulary[24], as none was yet foreseen. The proposal derives from the ELI ontology[25]. This is done in the expectation that a webmaster can use it to disseminate more structured data about legislation and also in the hope that search engines will use such data for providing added value to their users. This *could* lead to improved search results as depicted in the mock-up in Figure 1, where metadata of a given act are directly shown in the search result: the title, the status, domain keywords and direct access to the original or latest consolidated version.



EUR-Lex – 31980L0181 – EN – EUR-Lex

Council Directive of 20 December 1979 on the approximation of the laws of the Member States relating to units…

Status : Currently in force About : metrology, measuring equipment, approximation of laws

Original version (20/12/1979) | latest consolidation (27/05/2009)

Figure 1. Mock up of improved search results

This schema.org legislation extension has already been implemented by Luxembourg and the Brazilian parliament[26].

[21] http://orbilu.uni.lu/handle/10993/31825.

[22] http://schema.org.

[23] ELI Taskforce mandate: https://eur-lex.europa.eu/content/eli-register/governance_rules.pdf.

[24] https://pending.schema.org/Legislation.

[25] Interested readers can follow the archive of the discussion in the introduction of this extension in schema.org at: https://github.com/schemaorg/schemaorg/issues/1156.

[26] https://github.com/schemaorg/schemaorg/issues/1743#issuecomment-438768067.

4.3. Analysis and Navigation in Legal Knowledge Graphs

The availability of interconnected legislation information will allow to search across the global data graph and provide answers to queries such as “how is this EU Directive implemented in different countries?”, “where can I find a translation of French legislation into English?”, or “what legislation exists about ‘food inspection’ and ‘labelling’ across EU Member States?”. We can also envisage a fine-grained transposition analysis, with e.g. side-by-side display of directive articles and how they are transposed in a national corpus.

This will also improve navigation inside a legal act; a proof of concept for this functionality is the Lexparency prototype[27]. A related data visualization – not based on ELI – to navigate the GDPR was made at the CNIL - Commission National de l’Informatique et des Libertés[28].

The availability of reusable content formats and machine-readable links in a global legal knowledge graph on the World Wide Web can also serve as a basis for Artificial Intelligence algorithms to take legislation information into account.

4.4. Description of Early Legislation Drafts

As a legal watchman, how can I be notified that a legislative project has been issued, which could have a potential impact on the domain that I am monitoring? How can I access the impact studies or debate recordings of the legislative project that preceded a piece of legislation, and how can I reuse this information? As a civil servant in an EU Member State, how can I be informed at an early stage that a directive might potentially be amended, in order to allow me to start preparing for national transposition? These scenarios require not only machine-readable descriptions of legislation but also machine-readable descriptions of draft legislation at an early stage.

To address these use-cases, ELI has released ELI for Draft Legislation (ELI-DL)[29], an extension of the core ELI ontology to add structured metadata in the description pages of legislative projects. This extension emphasizes the description of *events* occurring during the legislative workflow, as a complement to the description of the sole *documents*. As of March 2019, this non-official draft ontology extension is open for comments.

4.5. Exchange of Data Between Legal Information Systems

A higher degree of interoperability between legal information systems is one of the central goals of ELI. This is particularly interesting in a European institutional context where data is exchanged between the European Commission and EU Member States in the frame of the transposition notification mechanism. ELI – and Web of data technologies in general – enables this data exchange in two ways: when starting a transposition project, EU Member States can fetch the ELI metadata of the Directive to be transposed from the EUR-Lex site managed by the Publications Office of the EU and integrate it in their own legal information system. Secondly, when a Member State wants to notify the Commission that the transposition is effective, it can point to the ELI reference(s) of the trans-

[27] https://lexparency.org/.

[28] https://www.cnil.fr/fr/reglement-europeen-protection-donnees/dataviz.

[29] https://joinup.ec.europa.eu/solution/eli-ontology-draft-legislation-eli-dl.

posing act(s), so that the Commission integrates the metadata of the transposed text in its own database. It is interesting to outline that this happens without any proprietary service or protocol, using only web standards.

4.6. Addressing Multilingual and decentralized Aspects

ELI allows to efficiently inter-link legislation in multilingual and/or federal legal environments. The FRBR hierarchy makes it possible to have a legal resource described only once associated to multiple translations. Links in the ontology allow to interrelate resources published in national and regional OJs, like in the Spanish implementation of ELI[30].

4.7. Linking with Case Law and Normative Requirements

Other potentialities of ELI lie in the ability to create links to and from other data sources, such as the European Case Law Identifier (ECLI)[31] or from the analysis of the normative requirements encoded in the legal provisions of the act [3].

5. ELI in the Context of Legal Knowledge Graphs

5.1. Building a 'Legal Knowledge Graph'

ELI, in line with the approach of the Web of data, with an FRBR-based data structure, encourages legal publishers to move from a document-centric perspective toward a *data-centric perspective*. In this approach, access to content (here, the legislation text) is enabled by a database containing the structured description of all the entities comprised in the knowledge domain. In the case of a legal publisher, such entities are the interrelated FRBR levels of the content notices, as well as the supporting concepts for the description of the notices: types of acts, thematic keywords, etc. Content files in multiple formats (XML, HTML, PDF) will also be stored in it. Such a database of highly interconnected entities plays a central role in the information system, as it can be reused across many applications. It is often referred to as 'knowledge graph', and in the case of legal publishers as 'legal knowledge graph'.

We describe in this Section the benefits we see in this style of architecture, exemplified in the information systems of Luxembourg and the Publications Office of the EU.

5.2. Characteristics of a Legal Knowledge Graph

The very nature of a graph data structure makes it easy to *aggregate data from heterogeneous data sources*, and such a legal knowledge graph links the data from multiple business applications: the Official Journal, the management of directive transpositions, consolidations, legislation projects, archives, international treaties, etc.; applications can navigate the links to operate transversally across data silos. The graph is controlled by an *ontology* and consolidates existing values in each application through *controlled vocabu-*

[30] http://administracionelectronica.gob.es/ctt/eli/descargas.

[31] https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-en.do.

laries shared across the whole system that serve as a ‘pivot’ to link the data. The graph is seen as a Data Warehouse and becomes the *single source of truth*[32] for the dissemination of data through all channels: web portals, ELI metadata, open data portals, APIs, RSS feeds, etc. As the system evolves, it tends to become also the single source of truth *inside* the information system, and not only for the dissemination to the outside world.

5.3. Legilux Casemates and the OP Cellar: Two Legal Knowledge Graphs in Action

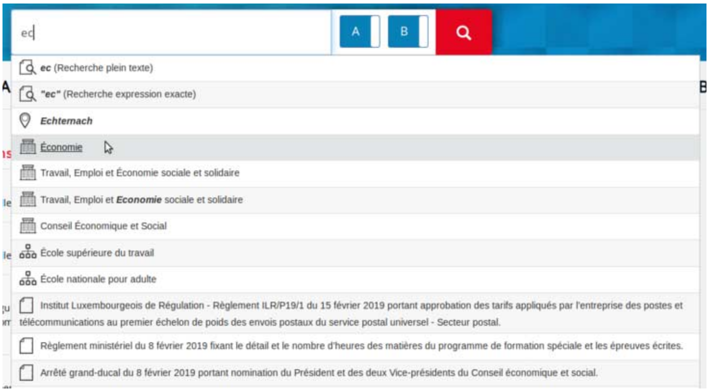
The Publications Office of the EU has its legal knowledge graph stored in the Cellar[33], its semantic repository containing structured description of all EU legislation. The Cellar is the source of data for the EUR-Lex portal[34], including the embedded ELI metadata. The Cellar database can be freely queried from the outside[35].

Legilux[36], the web portal of the Official Journal of the Grand Duchy of Luxembourg, also relies on a legal knowledge graph, Casemates[37].

We describe here some examples of accessibility features permitted by such knowledge graph architectures.

An obvious feature is that they provide a *single web portal to access all the content*: the different OJ series, legislative projects, international agreements, treaties and more.

These portals provide classical *faceted search* features on their content, based on the metadata of each act and the associated controlled vocabularies. Legilux even provides a *semantic auto-complete* feature for its search input field (Figure 2): not a plain search field, not proposals based on frequent queries, but on entities of the knowledge domain; as users type words in the search field, they are prompted with names of ministries, keywords, places and titles of laws corresponding to the letters typed, thus enabling quicker access to content.



ec

ec (Recherche plein texte)

"ec" (Recherche expression exacte)

Echternach

Économie

Travail, Emploi et Économie sociale et solidaire

Travail, Emploi et ***Economie*** sociale et solidaire

Conseil Économique et Social

École supérieure du travail

École nationale pour adulte

Institut Luxembourgeois de Régulation - Règlement ILR/P19/1 du 15 février 2019 portant approbation des tarifs appliqués par l'entreprise des postes et télécommunications au premier échelon de poids des envois postaux du service postal universel - Secteur postal.

Règlement ministériel du 8 février 2019 fixant le détail et le nombre d'heures des matières du programme de formation spéciale et les épreuves écrites.

Arrêté grand-ducal du 8 février 2019 portant nomination du Président et des deux Vice-présidents du Conseil économique et social.

Figure 2. Legilux: Semantic auto-complete feature for search input

[32] https://en.wikipedia.org/wiki/Singlesourceof_truth.

[33] Cellar guide for data reusers: https://publications.europa.eu/en/publication-detail/-/publication/50ecce27-857e-11e8-ac6a-01aa75ed71a1/language-en/format-PDF/source-73059305.

[34] EUR-Lex: https://eur-lex.europa.eu/.

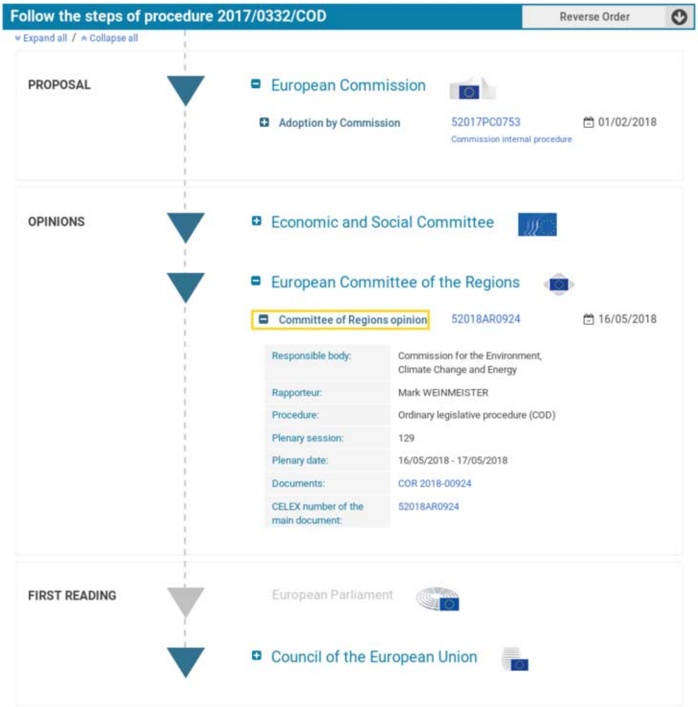
[35] Cellar SPARQL service: http://publications.europa.eu/webapi/rdf/sparql.

[36] Legilux: http://legilux.public.lu/.

[37] Casemates: http://data.legilux.public.lu/.

When users have accessed the notice of the legislation act they are interested in, the graph nature of the information enables to ‘*navigate anywhere*’: not only can they click on any of the acts somehow linked from or linking to this one, but they can also search for legislation with similar characteristics, by listing acts with the same author or thematic keywords.

Time and versioning of legislation is an important aspect of how the legal knowledge graph can be exploited, particularly in two situations: to show and navigate through the successive versions of an act, or to display the timeline of events in a legislative project. These display types require the aggregation of a lot of information from different entities. EUR-Lex offers a good example of such a timeline, where each event in a procedure can be opened for details, giving access to the documents linked to this step (Figure 3).



Follow the steps of procedure 2017/0332/COD

PROPOSAL

European Commission

Adoption by Commission 52017PC0753 Commission internal procedure 01/02/2018

OPINIONS

Economic and Social Committee

European Committee of the Regions

Committee of Regions opinion 52018AR0924 16/05/2018

Responsible body:	Commission for the Environment, Climate Change and Energy
Rapporteur:	Mark WEINMEISTER
Procedure:	Ordinary legislative procedure (COD)
Plenary session:	129
Plenary date:	16/05/2018 - 17/05/2018
Documents:	COR 2018-00924
CELEX number of the main document:	52018AR0924

FIRST READING

European Parliament

Council of the European Union

Figure 3. EUR-Lex: timeline of events in a legislative project

The knowledge graph can be queried transversally, based on any of its links and as such serves as a basis for flexible *data visualizations*. Legilux provides, for example, a quick access to the current transposition projects by ministry, type of text, and organisation[38].

Such legal knowledge graphs with a sufficient level of precision can allow for *automated consolidations* of legislation. This requires an unambiguous identification of legislation subdivisions, a precise annotation of legal references, and the ability to express the semantic links between amendments and original texts.

[38] Legilux access to transpositions: http://legilux.public.lu/data-graphics/transpositions.

The legal knowledge graph, as the data source of every dissemination channel, also serves for the dissemination of legislation datasets on open data portals, including the source content files, content metadata and supporting vocabularies used to index legislation.

6. Conclusion

The description of the European Legislation Identifier and the legal knowledge graph architectures shows the added value of this innovative approach for the benefit of legal information distribution. We outlined its potentialities as well as the challenges it will be facing in the coming steps. ELI is sufficiently generic to be adopted outside Official Journals and EU Member States and with its ‘benevolent’ approach, encourages both its implementation by legal information providers and the creation of innovative legal services by data reusers.

References

[1] Palmirani, M., Sartor, G., Rubino, R., Boer, A., de Maat, E., Vitali, F. & Francesconi, E. (2007). *Deliverable 3.2: Guidelines for Applying the New Format*. ESTRELLA European Project for Standardized Transparent Representations in Order to Extend Legal Accessibility.

[2] Mathis, B. (2019). La normalisation des documents normatifs, une avancée européenne au service de l’innovation juridique. In Masson, A. & Bouthinon-Dumas, H. (Eds.). *L’innovation juridique et judiciaire*. Larcier, 147-152.

[3] Gandon, F., Governatori, G. & Villata, S. (2017). Normative Requirements As Linked Data. In Wyner, A. & Casini, G. (Eds.). *Legal Knowledge and Information Systems. Proceedings of the 30th Jurix Conference*. IOS Press, 1-10.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190017

Improving Public Access to Legislation Through Legal Citations Detection: The Linkoln Project at the Italian Senate

Lorenzo BACCI [a], Tommaso AGNOLONI [a]
Carlo MARCHETTI [b] and Roberto BATTISTONI [b]
[a] *Istituto di Teoria e Tecniche dell'Informazione Giuridica – ITTIG-CNR (Italy)*
[b] *Senato della Repubblica (Italy)*

Abstract. In this paper we describe Linkoln, an open framework for the automatic detection and linking of legal references contained in legal texts. The problem was tackled by providing a modular and extensible approach in order to efficiently cover the wide variability and specific peculiarities of legal citation practices. The project was initiated in collaboration with the Italian Senate with the aim to make available to Italian legislative authorities and official publishing bodies, a robust and extensible automatic tool to improve access to published legislation. The result of this effort is Linkoln which was recently successfully integrated in the application serving documents on the institutional website of the Italian Senate to activate hyperlinks to cited legislation.

Keywords. natural language processing, legislative acts, legal references, linked open data

1. Background and Motivation

Hyperlinks to cited norms, preferably resolved at the provision granularity level, are essential to improve readability of a legislative text. Other than for enhancing the navigation of legislation, the extraction and annotation of machine readable legal references as metadata of legislative texts is essential to guarantee interoperability and to enable higher level applications. In the Legal Semantic Web and Linked Data perspective, the graph of legal references in the legislative corpus, with resources (nodes) identified with standard identifiers, can be exploited by applications for several use cases. In Legal Information Retrieval, machine readable information on the incoming and outgoing legal references of a document can be used to provide users with additional searching capabilities and to improve search results [1].

Manual reference tagging – sometimes offered by commercial legal publishers – is a labor intensive procedure not viable in the public domain, especially to cope with legacy archives and with the growing amount of documents published in national legal databases. On the other hand, despite the fact that drafting rules and citation guidelines exist at both national and EU level, exceptions to the recommendations are very frequent and the automatic legal reference extraction task faces the complexity of coping with a diversity of styles, variants and formats [2]; [3]; [4]; [5].

In 2015, after more than a decade from the introduction of persistent identifiers (*urn:nir*) [6] for Italian legislative sources, and from the release of *xmLegesLinker*[1] [7], the software developed by ITTIG-CNR for legislative reference parsing still in use in all major institutional websites of Italian legislative authorities, the Italian Senate promoted the complete redesign of the reference parsing software with the objective of improving its coverage, maintainability, usability at scale and ease of integration.

A wide network of interested actors – institutional stakeholders, issuing authorities and official publishing bodies – have been involved in the sharing of the objectives and the gathering of requirements, with the aim of favouring the adoption, reuse and iterative improvement of the tool. Within this collaboration ITTIG-CNR designed and developed Linkoln, the new software project for the automatic extraction of legal references from Italian legal texts. Linkoln is written in Java and released as open source software. Its first public release is the integration in ShowDoc, the application for the visualization of official acts on the public website of the Italian Senate.

2. Automatic Legal References Extraction With Linkoln

2.1. Terminology

In order to handle the task of identification and extraction of legal references from text, it is important to state the main concepts that are at play in this particular context.

We define a *citation* the fragment of text that is used by an author to refer to a document or to a partition of a document; a *textual feature* is an arbitrary sub fragment of text that conveys a specific meaning, or *entity*, within the citation (e.g. the authority that issued the referred document); a textual feature can be conceptually *atomic* (e.g. the month when the document was issued) or *complex*, i.e. composed by atomic and complex features (e.g. the date the document was issued, or an authority of a specific geographic location); when a textual feature is traced back to a list of possible known values or its *value* can be expressed in a standard format through a process of normalization, it becomes a *feature of the reference*; in this context, we call a *reference* the set of features that derives from a citation.

Only when a citation is identified, normalized and expressed as a reference it can be fully exploited by machines, both for further enrichment of the reference and linking (see section 3) and for indexing and querying a corpus on the documents that are cited in each text.

2.2. The Problem

The principal and most obvious hardship that need to be overcome by a software that performs automatic extraction of legal references is the one represented by the great amount of textual variants in which every possible value of every feature can be expressed or has been expressed in the past, and by the different patterns in which features can appear and be arranged in order to form a reference.

Besides that, further sources of complexity are constituted by *multiple references* and *aliases*. Multiple references derive from citations containing multiple partitions of

[1] www.xmleges.org.

the same legislative documents or from citations containing multiple documents that share a feature that is expressed only once, like a common issuing authority or a type of document (e.g. a list of laws). Aliases can be found in citations that, instead of referring to a document through its specific features, use a single textual expression through which the document is commonly known (e.g. European treaties, national codes, etc.).

In addition to identification and coverage issues, a software of this kind is required to be particularly efficient in doing its job, so that it can be integrated in pre-existing application or in web services in order to perform on-the-fly extraction and mark-up of references avoiding negative impacts on the user experience.

2.3. The Solution

Linkoln is the outcome of years of experience and work done at ITTIG-CNR on the topic of legal link extraction from texts written in Italian. Its design choices can be summarized as follows:

- the general approach to textual analysis is rule-based: to be able to have complete control at every step of the analysis is worth the effort of writing rules, listing values and patterns, especially in terms of overall accuracy of the software and ease of extension to new cases;
- the analysis is performed in several iterations of textual scanning, each time identifying more and more complex entities, based on the results of the previous iterations;
- the software architecture is basically a pipeline of services, where each service accomplishes one iteration of the analysis and transmits the results to the next service that uses it as its input and so on;
- an internal annotation system is introduced for storing (and reading) the results of each service directly within the text that is passed on along the pipeline;
- every annotation is linked to a hierarchy of in-memory Java objects;
- for the implementation of extremely efficient textual scanners within each service, Linkoln uses JFlex[2], a Java tool which, through the definition of regular expressions and macros of regular expressions, rules and states (or start conditions), allows the development of deterministic finite automata for text analysis that can be exploited in a Java environment.

This kind of textual analysis, achieved through a number of modules that implement the various services of the pipeline, brings several benefits, since:

- a clear separation of codes and rules is of crucial importance when the rules are so many and entities and patterns are so intertwined;
- it is extremely hard to identify and normalize complex entities with a single set of regular expressions or a single scanner or automata in an effective way: Linkoln starts from atomic entities and then builds on them more and more complex recognitions in an arbitrary order.

Having both an annotation system and in-memory objects corresponding to every single annotation, linked through an *id*, is also extremely beneficial:

- the annotation system within the original text allows to perform the processing of each service completely as a textual analysis based on regular expressions, states, etc., exploiting the efficiency of JFlex to the fullest;

[2] www.jflex.de.

- in-memory objects linked to every annotation through an *id* make possible to store as much additional info as we want concerning a specific annotation (besides the type of entity and the value that are also stored in the annotation).

2.4. The Annotation System

The basic structure of the internal annotations introduced by Linkoln services is:

```
[LKN:NAME:VALUE:ID] text [/LKN]
```

While LKN is a constant and acts as a namespace, NAME indicates the type of entity that is being annotated in a normalized form (e.g. "EU_LEG_AUTH" for European legislative authorities), VALUE is the normalized value for such entity (e.g. "EU_COUNCIL" for the European Council) and ID is the numeric identifier that links the specific annotation to an in-memory object (e.g. an instance of the class *European-LegislationAuthority*, which extends *Authority*, which extends *AnnotationEntity*).

2.5. Linkoln in Action

Starting from a citation to a specific partition of a decree issued by the Ministry of Finance, such as *"..vedi lett. e), comma 2,f art. 2 del decreto del Ministero delle finanze del 25 novembre 1998, n. 418..."*, the pipeline of services is put in motion. One service is specialized in the recognition of Italian Ministries:

```
vedi lett. e), comma 2, art. 2 del decreto del
[LKN:LEG_AUTH:IT_MIN_FINANZE:13]Ministero delle finanze[/LKN]
del 25 novembre 1998, n. 418
```

and one after the other, many services do their annotation job, also instantiating the corresponding in-memory objects:

```
vedi [LKN:LETTER:E:3]lett. e)[/LKN], [LKN:COMMA:2:2]comma 2[/LKN],
[LKN:ARTICLE:2:1]art. 2[/LKN] del [LKN:LEG_DOCTYPE:MIN_DECREE:14]
decreto[/LKN] del [LKN:LEG_AUTH:IT_MIN_FINANZE:13]Ministero delle
finanze[/LKN] del [LKN:DATE:1998-11-25:11]25 novembre 1998[/LKN],
[LKN:NUMBER:418:12]n. 418[/LKN]
```

A complex kind of entity that can be found in legal references is the partition. Correctly identifying and annotating it is not trivial, especially when dealing with multiple partitions. Linkoln does it with a specific service that exploits the previous annotations of atomic entities such articles, commas and letters:

```
vedi [LKN:LEG_PARTITION:ART-2_COM-2_LET-E:16]lett. e),
comma 2, art. 2[/LKN] del...
```

A specific service is also used to recognize patterns of features that form a reference, by merging them in a legislative reference annotation:

```
...[LKN:LEG_REF::15]decreto del Ministero delle finanze
del 25 novembre 1998, n. 418[/LKN]
```

Note that the corresponding *LegislativeReference*† object (with *id* “15”) is linked to every object that corresponded to the annotations of the merged entities, so no information is lost. Finally, another service associates partitions and references:

```
vedi [LKN:LEG_REF::15]lett. e), comma 2, art. 2 del decreto del
Ministero delle finanze del 25 novembre 1998, n. 418[/LKN]
```

The original text is now enriched with the annotations of the references found by Linkoln and in-memory objects hold every information about them.

3. Identifiers and Resolution

The final stage required to access the cited norms from their textual citations, is the translation of the recognized legal reference into a standard identifier and eventually in a resolvable link. Standard identifiers for the legal domain have been introduced to allow the unique identification of legal sources as resources on the Web, and therefore to favour interoperability and interlinking.

3.1. urn:nir

The *urn:lex* standard [6] and its Italian implementation *urn:nir* define the scheme of a globally unique, persistent, location independent and language-neutral identifier. The *urn:nir* is a *transparent* identifier, i.e. an identifier that can also be read and composed by humans. It is in fact expressed by a serialization of the same features commonly used to cite a legal act:

DECRETO LEGISLATIVO 24 febbraio 1998, n. 58
`urn:nir:stato:decreto.legislativo:1998-02-24;58`

For Italian legislative sources *urn:nir* is the officially recommended identification standard. *urn:nir* is supported by Normattiva[3], the Italian portal of in-force legislation published by the Official Gazette. By adding the address of the *urn* resolver[4] as a prefix to the identifier, it is possible to obtain an *url* where a manifestation of the referred document is published. The *urn:nir* resolver in Normattiva also supports pinpoint of the specific version of the document in-force at a given date and resolution to the level of partitions.

3.2. CELEX

Similarly, for European legislative sources (Treaties, Directives and Regulations), the CELEX number is the unique identifier assigned – regardless of the language they are written in – to all documents published by the EUR-Lex portal of the Publications Office of the European Union. The CELEX number is a transparent identifier, as well. It is composed of four parts (*Sector, Year, Doc Type, Doc Nr.*)[5]. CELEX uses an encoding for the *Sector* (3 is for Legislative Acts) and for the *Document Type* (L is for Directives) and the same numbering used in textual citations on four digits:

[3] https://www.normattiva.it.
[4] https://www.normattiva.it/uri-res/N2Ls?
[5] See [8] for the detailed specification.

Direttiva 2000/60/CE del Parlamento europeo e del Consiglio, del 23 ottobre 2000
`CELEX: 3 2000 L 0060`

From the CELEX number, the *url* linking to a manifestation in a specific language or in a specific format, can be easily composed:

`https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:`

3.3. ELI

More recently, the Council of the European Union with the adoption of the *Council conclusions inviting the introduction of the European Legislation Identifier (ELI)*[6], pushed for the integration of European and national Official Gazettes and for the use of Linked Data technologies as a mean for resources integration. The European Legislation Identifier (ELI) provides a solution to uniquely identify and access national and European legislation online. It includes:

- Web identifiers (HTTP-URIs) for legal sources;
- metadata specifying how to describe legal sources according to the underlying ELI Ontology;
- recommendation to participating Member States to embed such metadata elements in machine-readable formats into the webpages of their Legal Information Systems.

The implementation of the ELI identification Scheme is based on URI templates[7] specific to national jurisdictions. Each Member State will build its own, self-describing URIs using the described components as well as taking into account their specific language requirements. ELI has been recently implemented both at the EU and Italian level and in several other Member States[8].

The implementation of ELI for European legislation relies on a syntax based on the composition of the reference features:

`/eli/typedoc/year/naturalnumber/oj`

The EU Publications Office implements URI templates for all acts published on the Official Journal, series L (Legislation). Currently Directives (`dir`), Regulations (`reg`) and Decisions (`dec`) have been fully implemented[9].

Direttiva 2000/60/CE del Parlamento europeo e del Consiglio, del 23 ottobre 2000
`https://data.europa.eu/eli/dir/2000/60/oj.`

The Italian implementation of ELI by the Italian Official Gazette[10] uses the following URI template for the identification of legal acts:

`/eli/id/yyyy/mm/dd/codiceRedazionale/tipoSerie`

where the full date is the date of publication in the Official Gazette and the field `codiceRedazionale` is the unique identifier of the legal act in the Official Gazette, an alphanumeric code which is typically unknown to users and never used in textual citations. For example:

DECRETO LEGISLATIVO 24 febbraio 1998, n. 58
`https://www.gazzettaufficiale.it/eli/id/1998/03/26/098G0073/sg`

[6] Council conclusions inviting the introduction of the European Legislation Identifier (ELI) 2012/C 325/02, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XG1026(01).

[7] https://tools.ietf.org/html/rfc6570.

[8] https://eur-lex.europa.eu/eli-register/implementation.html.

[9] https://eur-lex.europa.eu/eli-register/eu_publications_office.html.

[10] https://eur-lex.europa.eu/eli-register/italy.html.

3.4. Identifier Generation

The translation of the legal reference into a standard identifier therefore depends on how the identifier is designed. For *transparent* identifiers, once known the reference features and the composition rules, the identifier can be programmatically generated without recourse to external knowledge. This is the case of *urn:nir* for Italian legal acts and CELEX and ELI-EU for European legislation, for which Linkoln provides support with dedicated identifier generation services.

If the standard identifier contains features which are different or typically not used in textual citations, e.g. a different numbering system, the reference features extracted from textual citations cannot be used or are not sufficient to directly compose the identifier. In such a case, a register of the identifiers of all the existing legal acts should exist, made available for programmatic access and kept up-to-date as new legislative measures are enacted.

A data based identifier generation strategy will therefore require to query the register of identifiers based on the reference features in order to match the corresponding identifier.

For the generation of ELI for Italian legislative acts, such data can be obtained from:

1. An RDFa crawler[11] ran on the pages of the website *gazzettaufficiale.it*;
2. The bulk dataset of Italian ELIs in RDF published on the European Open Data portal[12] (which might suffer a delayed update for recently enacted norms);
3. A richer dataset obtained from the scraping of the HTML pages of the Italian Official Gazette.

In Linkoln, data obtained from the above official sources, represented according to the ELI RDF metadata model[13], are put behind a SPARQL endpoint exposing the up-to-date dataset. A data based identifier generation service, queries such data with features extracted from the textual citation in order to match the corresponding identifier. Currently ELI metadata suitable for identification embedded in Italian legal sources published by the Official Gazette include: *issuing date*, *publication date*, *document type*[14] and *issuing authority*[15]. Document numbers (for numbered acts) are not made available.

Therefore a query to the register in its current stage of development can still give more than one match that cannot be disambiguated given the available information. Eventually, the dataset of Italian ELI resources can be enriched with additional triples obtained from non-official sources.

Linkoln provides support for all the aforementioned identification standards either through specialized services for their composition or through services for matching references metadata with external datasets. Access to additional reference resolution services (e.g. provided by commercial publishers) can be configured based on specific needs.

[11] http://labs.sparna.fr/eli-validator/documentation#howto-crawler.

[12] https://data.europa.eu/euodp/en/data/dataset/eli-european-legislation-identifier-italy.

[13] https://publications.europa.eu/en/web/eu-vocabularies/eli.

[14] https://www.gazzettaufficiale.it/eli/tables/resource-type.

[15] https://www.gazzettaufficiale.it/eli/tables/issuers.

4. Input, Output and Integration in ShowDoc

Besides plain text, Linkoln supports previously annotated texts as input (HTML, XML) and is able to render the additional reference annotation while preserving pre-existing annotations. From the internal annotation performed along the pipeline, several outputs can be obtained and configured based on specific needs: HTML rendering of the annotated text for hyperlinks navigation, XML annotation of references according to a specific Schema, raw structured data exposed e.g. as RDF or JSON for storage or subsequent consumption by applications. Distributed as a Java Library with a standard Java API, the software can be easily integrated within existing web or standalone applications or be wrapped into an interoperable HTTP API for integration with external distributed services. It can either be used for automatic batch parsing of a large corpus or invoked by a remote user-interface over text fragments or entire documents. Especially for this latter use case, the overall efficiency of the software guarantees a quick response (in terms of user experience) even with large texts as input.

This capability is exploited in the integration with ShowDoc, the web application for the visualization of official acts (including legislative proposals, amendments, dossiers, etc.) on the website of the Italian Senate. The ShowDoc web application serves the XHTML versions of parliamentary documents to the users of Italian Senate web site since 2001. In order to improve user experience, the IT Department of the Italian Senate recently developed a new version of this web app, exploiting mainstream technologies, as well as consolidated web paradigms and technological architectures. During its development, Linkoln source code has been directly integrated into the ShowDoc backend in order to to minimize invocation times with respect to wrap Linkoln and exposing it as a ReST service. Thanks to the availability of the source code and to the overall design choices, Linkoln integration has been straightforward.

Among its functionalities ShowDoc allows users to detect legislative citations and activate hyperlinks while reading a document or document partition by invoking Linkoln on the currently visualized HTML fragment. In this scenario Linkoln receives previously marked-up HTML text as input and returns the enriched HTML annotation with <*a*> tags in correspondence with the detected legislative references (see an example in Figure 1). Text analysis is performed on-the-fly and the result rendered for visualization.

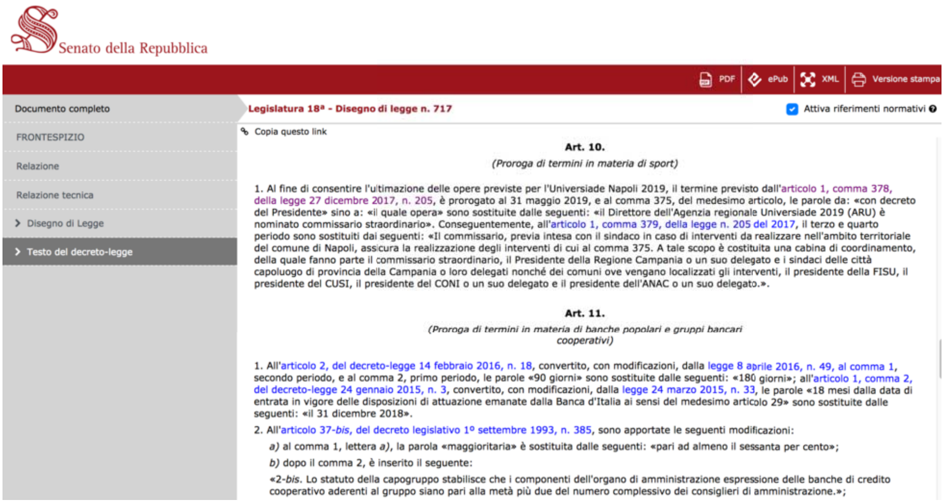
5. Conclusions and Future Work

The opportunity to reengineer from the foundations our legal citations detection software, resulted in a comprehensive framework for efficient rules based citation extraction from legal texts, distilling years of experience gained in the field.

Linkoln is released by ITTIG-CNR as open source software licensed with the General Public License GPL v.3. The code repository and a demo of the latest version for testing and evaluation purposes are publicly available[16].

In order to evaluate the accuracy and the coverage capabilities of the software, hundreds of citations have been collected from real-world legislative texts, including citations to European, national and regional legislation, involving several typologies of legislative documents, issued by a wide number of different authorities, with different sys-

[16] https://ittig.gitlab.io/linkoln-senato.



Art. 10.

(Proroga di termini in materia di sport)

1. Al fine di consentire l'ultimazione delle opere previste per l'Universiade Napoli 2019, il termine previsto dall'articolo 1, comma 378, della legge 27 dicembre 2017, n. 205, è prorogato al 31 maggio 2019, e al comma 375, del medesimo articolo, le parole da: «con decreto del Presidente» sino a: «il quale opera» sono sostituite dalle seguenti: «il Direttore dell'Agenzia regionale Universiade 2019 (ARU) è nominato commissario straordinario». Conseguentemente, all'articolo 1, comma 379, della legge n. 205 del 2017, il terzo e quarto periodo sono sostituiti dai seguenti: «Il commissario, previa intesa con il sindaco in caso di interventi da realizzare nell'ambito territoriale del comune di Napoli, assicura la realizzazione degli interventi di cui al comma 375. A tale scopo è costituita una cabina di coordinamento, della quale fanno parte il commissario straordinario, il Presidente della Regione Campania o un suo delegato e i sindaci delle città capoluogo di provincia della Campania o loro delegati nonché dei comuni ove vengano localizzati gli interventi, il presidente della FISU, il presidente del CUSI, il presidente del CONI o un suo delegato e il presidente dell'ANAC o un suo delegato.».

Art. 11.

(Proroga di termini in materia di banche popolari e gruppi bancari cooperativi)

1. All'articolo 2, del decreto-legge 14 febbraio 2016, n. 18, convertito, con modificazioni, dalla legge 8 aprile 2016, n. 49, al comma 1, secondo periodo, e al comma 2, primo periodo, le parole «90 giorni» sono sostituite dalle seguenti: «180 giorni»; all'articolo 1, comma 2, del decreto-legge 24 gennaio 2015, n. 3, convertito, con modificazioni, dalla legge 24 marzo 2015, n. 33, le parole «18 mesi dalla data di entrata in vigore delle disposizioni di attuazione emanate dalla Banca d'Italia ai sensi del medesimo articolo 29» sono sostituite dalle seguenti: «il 31 dicembre 2018».

2. All'articolo 37-*bis*, del decreto legislativo 1º settembre 1993, n. 385, sono apportate le seguenti modificazioni:

a) al comma 1, lettera *a)*, la parola «maggioritaria» è sostituita dalle seguenti: «pari ad almeno il sessanta per cento»;

b) dopo il comma 2, è inserito il seguente:

«2-*bis*. Lo statuto della capogruppo stabilisce che i componenti dell'organo di amministrazione espressione delle banche di credito cooperativo aderenti al gruppo siano pari alla metà più due del numero complessivo dei consiglieri di amministrazione.»;

Figure 1. A screenshot of ShowDoc with activated citations detection

tems for expressing numbers and dates. The list of selected samples includes citations to partitions of a document (paragraphs, letters and items), multiple citations to different partitions of the same document and multiple citations to different documents issued by the same authority. Moreover, citations expressed through aliases or common names, both at the European and national level, are included. On this benchmark the software shows an accuracy (i.e. correctly recognises the features that compose the reference) higher than 90%. An evaluation campaign of the automatic extraction and linking performance on a manually supervised test corpus of legislation varied temporally and by enacting authority is planned.

Thanks to its modular architecture and flexibility, Linkoln can be further developed in several directions and exploited as a building block for several additional use cases.

Straightforward extensions are the recognition and resolution of internal references and the adaptation to the peculiarities of texts enacted by specific authorities (for example Regional and local authorities). In the parliamentary and legislative process domain, extensions to the detection of citations to legislative proposals, commission reports and acts resulting from the legislative life cycle in the EU Parliament and Commission, are an objective at hand.

A parallel process, finally converged in the current version of Linkoln (2.0), has been successfully undertaken in the judicial domain with projects for the automatic detection of case law and legislative references from case law texts [4]; [9].

Similarly, the flexibility of the approach is demonstrated by the possibility of multilingual extensions to texts written in different languages or issued by foreign jurisdictions, extension put in practice in [4] and recently tested in a proof-of-concept for the EU Publications Office for the analysis of the parallel linguistic versions of European legislation published on the EUR-Lex portal.

The increasing availability of Open Data on Italian and European legislative sources can be exploited for additional use cases, e.g. validation of the extracted references, resolution of incomplete references, suggestion of canonical textual citations. Besides being a user for the Linked Data cloud of legal sources, Linkoln can also be part of its automated construction by extracting structured information from big textual corpora.

After the adoption and integration in the publication workflow of the Italian Senate and thanks to its compliance with both legal and web open standards and its release as open source software, several Italian authorities and Public Administrations, will be encouraged to test and adopt the software and contribute to its iterative refinement, evolution and maintenance.

References

[1] van Opijnen, M. & Santos, C. (2017). On the Concept of Relevance in Legal Information Retrieval. *Artificial Intelligence and Law, 25*(1), 65-87.

[2] van Opijnen, M., Verwer, N. & Meijer, J. (2015). Beyond the Experiment: The Extendable Legal Link Extractor. In *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL), https://ssrn.com/abstract=2626521.

[3] Mowbray, A., Chung, P. & Greenleaf, G. (2016). A Free Access, Automated Law Citator with International Scope: The LawCite Project. *European Journal of Law and Technology, 7*(3), http://ejlt.org/article/view/496/691.

[4] Agnoloni, T., Bacci, L. & van Opijnen, M. (2017). BO-ECLI Parser Engine: The Extensible European Solution for the Automatic Extraction of Legal Links. In *2nd Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, held in conjunction with the 2017 International Conference on Artificial Intelligence and Law (ICAIL), http://ceur-ws.org/Vol-2143/paper4.pdf.

[5] Rustad, K. & McDonald, R. (2012). *Building a Free, Open Source Legal Citator*, https://www.ischool.berkeley.edu/projects/2012/building-free-open-source-legal-citator.

[6] Spinosa, P. L., Francesconi, E. & Lupo, C. (2009-). A Uniform Resource Name (URN) Namespace for Sources of Law (LEX), https://datatracker.ietf.org/doc/draft-spinosa-urn-lex/.

[7] Francesconi, E., Marchetti, C., Pietramala, R. & Spinosa, P. L. (2010). URN-based Identification of Legal Acts: The Case of the Italian Senate. *Informatica e diritto*, (1-2), 233-252, http://www.ittig.cnr.it/EditoriaServizi/AttivitaEditoriale/InformaticaEDiritto/IeD2010_1-2_FrancesconiEtal.pdf.

[8] Publications Office of the EU (2017). *How CELEX Numbers Are Composed. Version 201704*, https://eur-lex.europa.eu/content/tools/HowCelexNumbersAreComposed.pdf.

[9] Bacci, L., Francesconi, E. & Sagri, M.-T. (2013). A Proposal for Introducing the ECLI Standard in the Italian Judicial Documentary System. In Ashley, K. D. (Ed.). *Frontiers in Artificial Intelligence and Applications. Proceedings of the 26th Conference*. IOS Press, 49-58.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190018

Akoma Ntoso for Making FAO Resolutions Accessible

Monica PALMIRANI
CIRSFID, Università degli Studi di Bologna (Italy)

Abstract. Akoma Ntoso is an international legal XML standard, whose technical specifications are now approved by the OASIS body. The standard has been developed to model legislative, parliamentary, and judicial documents using Semantic Web design principles. However, other types of normative and regulatory documents can benefit from being represented in Akoma Ntoso, making it possible to formally describe their structure, their components (e.g., attachments), their references to and from other documents, the semantic annotation of some peculiar parts of regulatory language (e.g., actions, purposes), the workflow of the creation process, and modifications over time. This Chapter presents a legal analysis of FAO Resolutions and how to apply Akoma Ntoso to interoperate with other UN documents (e.g., resolutions of the UN General Assembly). We also present the identifier naming convention for managing multilingual interconnection between documents (e.g., the UN manages six official languages). Finally, we present the ALLOT ontology application for improving semantic annotation in light of Linked Open Data. The combination of Akoma Ntoso and the ALLOT ontology makes it possible to enhance searching capacity and presentation accessibility.

Keywords. LegalXML, Akoma Ntoso, Semantic Web, ontology

1. Introduction

The Food and Agriculture Organization (FAO) of the United Nations is a technical agency of the United Nations that leads international efforts to eradicate hunger and malnutrition. FAO's governance[1] is composed of governing (such as the Conference and the Council) and statutory bodies. The Conference is the organization's body and meets once in every two years in regular session in which each Member Nation and Associate Member are represented. The biennial Conference approves resolutions in plenary sessions. Resolutions approved by the Conference are qualified as soft law (they are not binding on members) but belong in the space of the international law framework. Also, European institutions refer to FAO resolutions when they annex conventions. Thus, for example, Resolution 12/97 FAO, which includes the International Plant Protection Convention, is cited in European Council Decision 2004/597/EC[2]. The UN General Assembly cites FAO resolutions[3], as well as domestic law. The UN, as well as FAO, translates all

[1] *FAO Basic Texts*, http://www.fao.org/3/k8024e/k8024e.pdf.

[2] https://eur-lex.europa.eu/eli/dec/2004/597/oj.

[3] https://undocs.org/A/RES/72/72 "69. Takes note of resolution 9/2017, entitled "Observance of the International Day for the Fight against Illegal, Unreported and Unregulated Fishing", adopted by the Conference of the Food and Agriculture Organization of the United Nations at its fortieth session.

the documents in six official languages: English, French, Spanish, Arabic, Chinese, and Russian.

Resolutions adopted by the Conference are inserted in the Conference Reports (Figure 1).

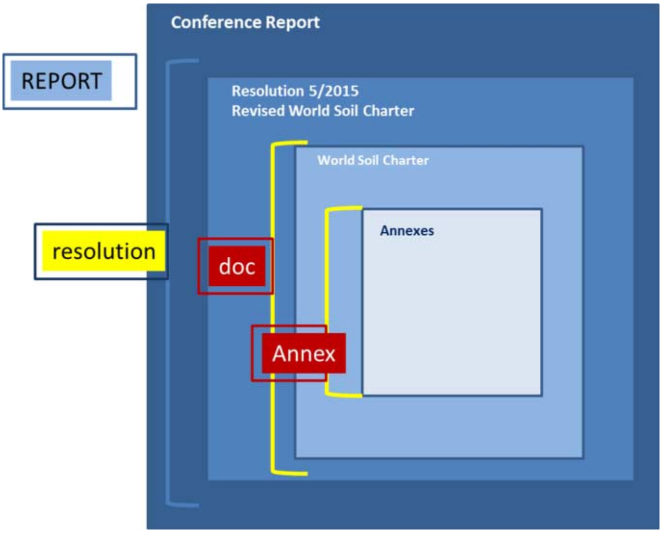


Figure 1. Modelling of a resolution within a Conference report

In the FAO portal resolutions are accessible using a separate list of web pages[4] published in HTML; the navigation system redirects to the Conference Report of the specific session where a given resolution is included. This organization of digital material has different problems: (1) end users have to browse a long page of Conference reports, including several resolutions, and can therefore loose track of their position while navigating; (2) the information system does not provide a tool for searching in the collection of resolutions, nor does it provide filters or channels (e.g., date, topic, keyword) for accessing the relevant material; (3) citations are not linked and navigable, so end users can't browse to another resolution cited in the Conference report; (4) switching between languages is managed at general level and not at document level; (5) there is no use of Semantic web classifications and tags; (6) it is not possible to download specific navigation results (e.g., a specific resolution in PDF or XML) – it is only possible to save the entire HTML.

End users in the current version of the front-office information system are not supported in navigation, and they need to know in advance which session the resolution was voted in. The back-office does not include an indexing system capable of effectively searching documents, aggregating them, and presenting them as a downloadable bulk. Finally, only the HTML format is available, and it is edited and formatted in different heterogeneous ways according to the historical period.

[4] http://www.fao.org/unfao/govbodies/gsbhome/conference/resolutions/2017/en/.

2. The AKN4UN Project

Akoma Ntoso [3] is an international XML standard approved by the OASIS standardization body. Even if Akoma Ntoso was designed for parliamentary, legislative, and judicial documents, it is also suited for soft-law documents like UN and FAO resolutions. The High-Level Committee on Management (HLCM), part of the UN System Chief Executives Board for Coordination, has set up a Working Group on Document Standards, and in April 2017 it adopted Akoma Ntoso as its document format[5]. This made it possible to define specific guidelines and customize the AKN schema for a UN scenario, and in particular for resolutions[6]. The AKN4UN documentation and schemas are the output of this analysis[7].

3. Akoma Ntoso for FAO Resolutions

3.1. Objectives

The FAO scenario was the first application of the AKN4UN strategic general plan. The aim of the FAO pilot project was to evaluate the use of Akoma Ntoso for FAO's documentation by improving back-office and front-office accessibility to resolutions. Using Akoma Ntoso the following objectives were achieved:

1. to switch documents from HTML to XML;
2. to use Semantic Web techniques to improve searching (e.g., using ontologies);
3. to interoperate with the different UN departments by way of citations;
4. to interchange the documents between different offices and members;
5. to simplify a document's lifecycle over the course of its drafting and management;
6. to track modifications over time, and track workflow;
7. to manage multilingualism;
8. to visualize document to improve transparency and accessibility with graph analysis of citations;
9. to make it possible to download search results in different formats (e.g., XML or PDF);
10. to design responsive interface for improving user experience.

3.2. Methodology

Co-design principles were applied as a basic methodology throughout the project, with periodical sessions in which requirements were analysed and defined, the tool was tested, and the metadata analysed. The analysis was the result of a deep legal, linguistic, and documental analysis of the requirements carried out with the main stakeholders of the FAO departments.

[5] https://www.unsystem.org/CEBPublicFiles/CEB-2017-3-HLCM33-Summay%20of%20Conclusions-FINAL.pdf.

[6] https://unsceb-hlcm.github.io/.

[7] textitGuidelines for the Mark-up of UN Normative, Parliamentary and Judicial Documents, High-Level Committee of Management, United Nations, https://unsceb-hlcm.github.io/.

The FAO staff marked up all the resolutions using the LIME web editor[8], customized for the FAO environment. Three members of the FAO staff received in-depth training, and they marked up 1,254 documents including the organization's Constitution. This makes it possible to establish a gold standard, deeply enriched with metadata and semantic annotation placed into the text coming from experts. The training was done by University of Bologna following the analysis done with FAO stakeholders.

The results are now available in a web portal for internal use, and thanks to XML Akoma Ntoso techniques, they include all the previously defined features. The functionalities make it possible to pursue all the previously mentioned objectives.

4. Digitalizing Resolutions, Eliciting Knowledge

After the stakeholder focus groups were formed and the interdisciplinary analysis was complete, we customized the LIME Editor to parse Microsoft Word documents of the original source of FAO resolutions. We used Regex and heuristics to maximize automatic markup and permit human experts to focus their time and expertise on the semantic part of the text's qualification. This made it possible to automatically detect the relevant parts of the text: preface, preamble, body, conclusions, annexes, references. We also used taxonomies and vocabularies for marking up words within sentences (e.g., *decides*, *noting*) and some relevant qualifications like SDG (Sustainable Development Goals).

Particular attention was devoted to marking up roles (e.g., Chair), persons, entities. Thanks to the ALLOT ontology [1], these annotations are stored in the <references> block of the Akoma Ntoso XML serialization, and the IRI to the appropriate ontology class is stored in the attribute @href.

The ALLOT[9] ontology is part of UNDO[10] (United Nations Document Ontology) [2], which makes it possible to specify context parameters through the TimeContextValue ontology[11]. Akoma Ntoso has seventeen Top Level Classes representing the basic pillars of reality: time, person, organization, concept, object, event, location, process, role, term, quantity, definition, entity. Thanks to UNDO, each of them is connected with values (e.g., a role), time (e.g., the period of time when a person occupies a role), and context (e.g., FAO). This makes it possible to carry out specific queries like *give me the documents during the time that Mr. XXX acted in the role of YYY, and not as ZZZ.* Figure 2 shows how the allot:Role class is a subclass of Value, which is related to tvc:withValue and tvc: atTime, capable of tracking time parameters. This mechanism makes it possible to attribute a role to some conditions, including a particular jurisdiction, as well as competences and powers that are valid in given time intervals.

5. Semantic Web Searches

The experts also qualified the parts of speech (PoS) using Akoma Ntoso term tags. Each term is connected with a type of action suggested by the sentence: for example, *decides*,

[8] http://bach.cirsfid.unibo.it/lime-fao/.

[9] https://w3id.org/akn/ontology/allot.

[10] https://unsceb-hlcm.github.io/onto-undo/.

[11] http://www.essepuntato.it/2012/04/tvc.

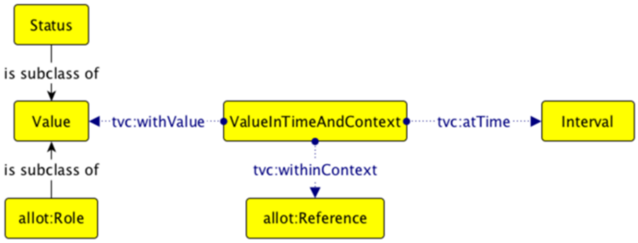
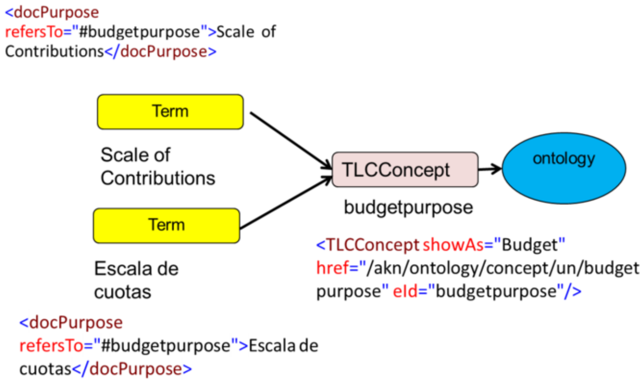


Figure 2. A TVC ontology pattern

authorises, or *requests* for operative sentences or *noting*, *having taken into consideration*, or *In recognition of* for preambular sentences. Using manual markup, we coordinated tokens of each term and we created a harmonized closed vocabulary *of actions*. We elected the best English token as a representative label for the class. All existing variants, including alternatives in the six official languages, are grouped under the main label. For example, *havingInMind* is a class that groups the following actions: *Having in mind* and *Having further in mind* and, in Spanish, the variant *Teniendo presente*. This is a method for creating a linguistic vocabulary starting from the terms present in the text so as to create concepts.

The same mechanism is applied to the document's purposes. Each resolution has a main purpose (e.g., food security and strategy). The experts marked up the text using the Akoma Ntoso inline element <docPurpose>. The attribution @*refersTo* is automatically detected in the text and trimmed (e.g., foodsecuritystrategypurpose). The different linguistic variants are harmonized with the main purpose categories that are expressed in the TLCConcept in the <reference> block (Figure 3).



```
<docPurpose
refersTo="#budgetpurpose">Scale of
Contributions</docPurpose>
```

```
<TLCConcept showAs="Budget"
href="/akn/ontology/concept/un/budget
purpose" eId="budgetpurpose"/>
```

```
<docPurpose
refersTo="#budgetpurpose">Escala de
cuotas</docPurpose>
```

Figure 3. The term, concept, ontology mechanism in Akoma Ntoso for FAO resolutions

Using roles, persons, organizations, purposes, and actions, we provide the end user with a very effective search tool. All these filters can be combined with structural elements (e.g., resolution number, date of adoption), temporal parameters (e.g., time interval for a search), and full-text retrieval in the text or in its content. Using the option 'search all in the same paragraph', the end user can restrict a search to requirements in the same provision.

Additionally, the auto-compose widget is used to support end users who do not have advance knowledge of the query they need to make. The search panel in Figure 4 enables end users to easily make very complex queries, exploiting Semantic Web annotation and the XML structure. For example, it is possible to make the following query: *Give me all resolutions that introduce amendments into Constitutions, but only in the time interval from 2007 to 2018*.

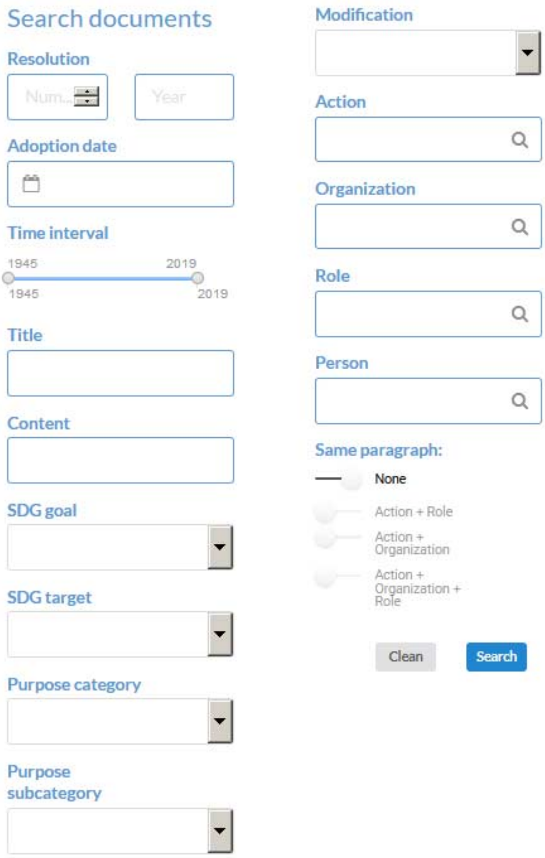


Figure 4. Search channels

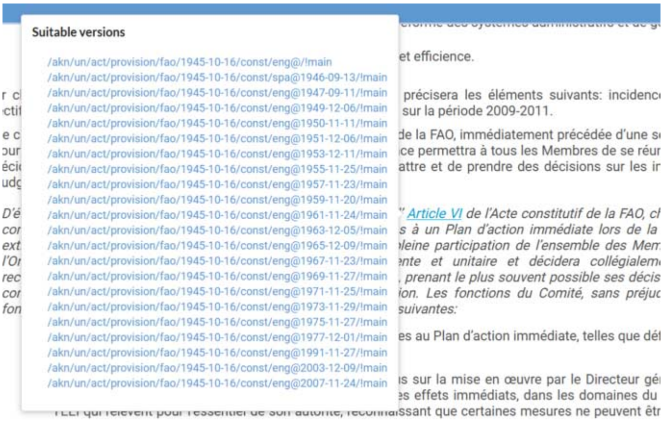
Finally we used the Sustainable Development Goals Interface Ontology (SDGIO)[12] to automatically classify resolutions when the text clearly mentions goals or the targets, as in this text fragment below: “Noting that camelids constitute the main means of subsistence for millions of poor families who live in the most hostile ecosystems on the planet, and contribute to the fight against hunger (SDG 2), the reduction of extreme poverty (SDG 1), the empowerment of women (SDG 5), and the sustainable use of terrestrial ecosystems (SDG 15)”.

This also applies to subgoals called targets: “Sustainable Development Goals (SDGs), particularly Target 2.5, related to genetic diversity”.

[12] https://github.com/SDG-InterfaceOntology/sdgio.

6. References and Navigation

The navigation of normative citations uses the Akoma Ntoso naming convention based on the FRBR ontology. Each citation is serialized by parsers and transformed into canonical IRIs. In dealing with cases where the the collection does not contain a corresponding document in the required language, we implemented cross-language navigation: in such cases, navigation automatically redirects the end user to the English version. Finally, we have point-in-time navigation. If a reference is connected to multiple versions, the portal presents all the admissible versions possible. In Figure 5 we can see that Article VI of the Constitution points to any version prior to the date when the current resolution was adopted (2007). End users can choose the best option according to their requirements.



Suitable versions

/akn/un/act/provision/fao/1945-10-16/const/eng@/!main
/akn/un/act/provision/fao/1945-10-16/const/spa@1946-09-13/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1947-09-11/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1949-12-06/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1950-11-11/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1951-12-06/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1953-12-11/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1955-11-25/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1957-11-23/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1959-11-20/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1961-11-24/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1963-12-05/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1965-12-09/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1967-11-23/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1969-11-27/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1971-11-25/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1973-11-29/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1975-11-27/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1977-12-01/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@1991-11-27/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@2003-12-09/!main
/akn/un/act/provision/fao/1945-10-16/const/eng@2007-11-24/!main

et efficience.

précisera les éléments suivants: incidenc
sur la période 2009-2011.

de la FAO, immédiatement précédée d'une s
ce permettra à tous les Membres de se réur
attre et de prendre des décisions sur les ir

" Article VI de l'Acte constitutif de la FAO, cl
s à un Plan d'action immédiate lors de la
leine participation de l'ensemble des Mem
ente et unitaire et décidera collégialem
, prenant le plus souvent possible ses décis
ion. Les fonctions du Comité, sans préjuc
suivantes:

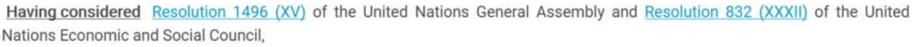
es au Plan d'action immédiate, telles que dé

s sur la mise en œuvre par le Directeur gé
es effets immédiats, dans les domaines du
ssant que certaines mesures ne peuvent êtr

Figure 5. Point-in-time navigation of references

7. Information Interchange

The use of Akoma Ntoso in FAO, and also in the UN General Assembly in New York, makes it possible to develop a model for information interchange. One of the best examples is with citations from a UN-New York resolution to FAO resolutions. In Figure 6 we can see the link to Resolution 1496 of the UN General Assembly and Resolution 832 of the UN Economic and Social Council:



Having considered Resolution 1496 (XV) of the United Nations General Assembly and Resolution 832 (XXXII) of the United Nations Economic and Social Council,

Figure 6. Citations to the UN General Assembly managed using the Akoma Ntoso naming convention

Following is an AKN XML fragment. The AKN resolver[13] can deference navigation to the appropriate portal according to the authority parameter (unga for the United Nations General Assembly, unecosoc for the United Nations Economic and Social Council):

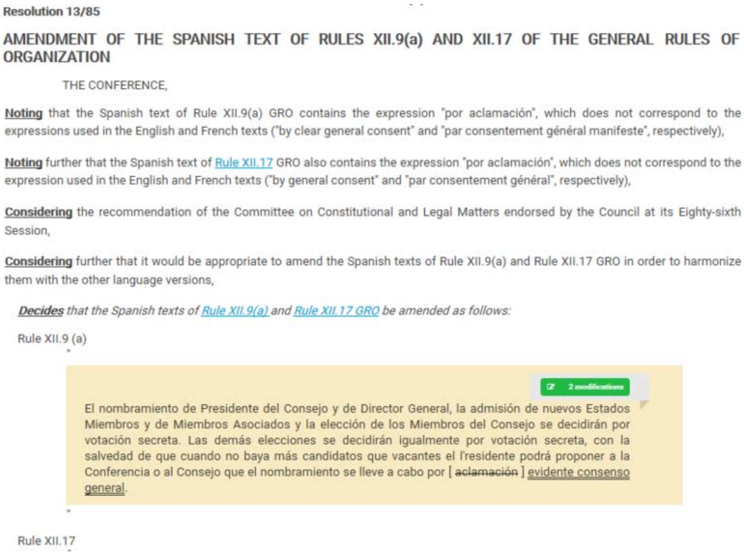
[13] http://akresolver.cs.unibo.it/.

```
<ref href="/akn/un/statement/deliberation/unga/1960-10-27/15-1496/!main"
eId="ref_1">Resolution 1496 (XV)</ref>
<ref href="/akn/un/statement/deliberation/unecosoc/1960-08-02/32-832/!main"
eId="ref_2">Resolution 832 (XXXII)</ref>
```

8. Modifications

Textual modifications are marked up in the text using the specific metadata <mod> in order to track each annotation in the text, also using <ins><del> to apply redline tracing. These modifications make it possible to follow the way resolutions evolve over time and to highlight these changes to decisionmakers.

End users wanting to see modifications can do so using a tab pop-up that recaps the information of the textual changes. Figures 7 and 8 show how Akoma Ntoso information can be used in the interface to make for good user experience.



Resolution 13/85

AMENDMENT OF THE SPANISH TEXT OF RULES XII.9(a) AND XII.17 OF THE GENERAL RULES OF ORGANIZATION

THE CONFERENCE,

Noting that the Spanish text of Rule XII.9(a) GRO contains the expression "por aclamación", which does not correspond to the expressions used in the English and French texts ("by clear general consent" and "par consentement général manifeste", respectively),

Noting further that the Spanish text of Rule XII.17 GRO also contains the expression "por aclamación", which does not correspond to the expression used in the English and French texts ("by general consent" and "par consentement général", respectively),

Considering the recommendation of the Committee on Constitutional and Legal Matters endorsed by the Council at its Eighty-sixth Session,

Considering further that it would be appropriate to amend the Spanish texts of Rule XII.9(a) and Rule XII.17 GRO in order to harmonize them with the other language versions,

Decides *that the Spanish texts of Rule XII.9(a) and Rule XII.17 GRO be amended as follows:*

Rule XII.9 (a)

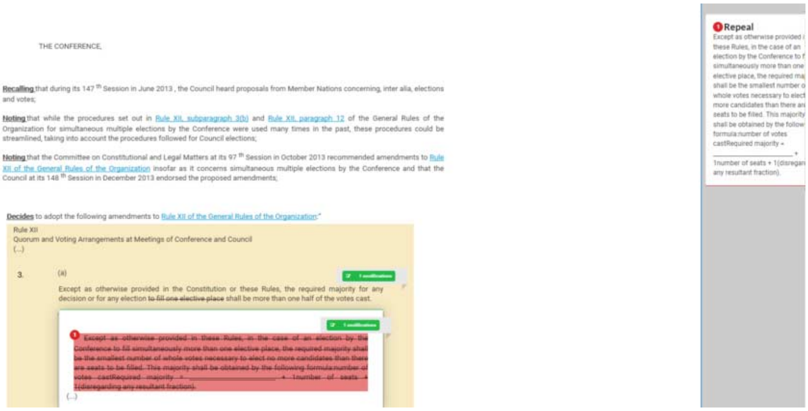
El nombramiento de Presidente del Consejo y de Director General, la admisión de nuevos Estados Miembros y de Miembros Asociados y la elección de los Miembros del Consejo se decidirán por votación secreta. Las demás elecciones se decidirán igualmente por votación secreta, con la salvedad de que cuando no baya más candidatos que vacantes el l'residente podrá proponer a la Conferencia o al Consejo que el nombramiento se lleve a cabo por [~~aclamación~~] evidente consenso general.

Rule XII.17

Figure 7. Modification made to the Spanish version only

9. Multilingualism

FAO's six official languages create a challenging environment in the form of multilingualism. The identifier naming convention for Akoma Ntoso elements makes it possible to synchronise the different linguistic variants and to develop special tools to help experts translate the text and not have to mark up the document anew. The Akoma Ntoso ID set is composed of a pair: eId (expression identifier) and wId (work identifier). In the Arabic version, the items in a list are lettered using specific Arabic alphabetical symbol. In order to synchronize them with the other alphabets, we use wId tracking the English master copy, and eId is the local language ID (Figure 9). This synchronization makes it possible



THE CONFERENCE,

Recalling that during its 147 th Session in June 2013 , the Council heard proposals from Member Nations concerning, inter alia, elections and votes;

Noting that while the procedures set out in Rule XII, subparagraph 3(b) and Rule XII, paragraph 12 of the General Rules of the Organization for simultaneous multiple elections by the Conference were used many times in the past, these procedures could be streamlined, taking into account the procedures followed for Council elections;

Noting that the Committee on Constitutional and Legal Matters at its 97 th Session in October 2013 recommended amendments to Rule XII of the General Rules of the Organization insofar as it concerns simultaneous multiple elections by the Conference and that the Council at its 148 th Session in December 2013 endorsed the proposed amendments;

Decides to adopt the following amendments to Rule XII of the General Rules of the Organization:"

Rule XII
Quorum and Voting Arrangements at Meetings of Conference and Council
(...)

3. (a)

Except as otherwise provided in the Constitution or these Rules, the required majority for any decision or for any election ~~to fill one elective place~~ shall be more than one half of the votes cast.

~~Except as otherwise provided in these Rules, in the case of an election by the Conference to fill simultaneously more than one elective place, the required majority shall be the smallest number of whole votes necessary to elect no more candidates than there are seats to be filled. This majority shall be obtained by the following formula:number of votes castRequired majority = ____________ + 1number of seats + 1(disregarding any resultant fraction).~~

(...)

Repeal
Except as otherwise provided
these Rules, in the case of an
election by the Conference to f
simultaneously more than one
elective place, the required ma
shall be the smallest number o
whole votes necessary to elect
more candidates than there ar
seats to be filled. This majority
shall be obtained by the follow
formula:number of votes
castRequired majority =
____________ +
1number of seats + 1(disregar
any resultant fraction).

Figure 8. A repeal within a resolution is modelled using the metadata relative to the deleted text

to manage cross-referencing links from one language to another; it enables browser citation during navigation; it also makes it possible to automatically detect the correct local eId from the text by following the local alphabet of the current language; and it enables multilingual searching (e.g., give me all points with eId="list_1__point_أ"):

<list eId="list_1"><point eId="list_1__point_أ" wId="list_1__point_a">
<num>(أ)</num><content eId="list_1__point_a__content">

Additionally, specific metadata are included in the XML AKN in order to connect the master copy with a translated version:

<FRBRtranslation fromLanguage="eng" by="#translator" authoritative="true"
href="/akn/un/statement/resolution/FAO/2011-07-02/13-2011/eng@/!main"/>

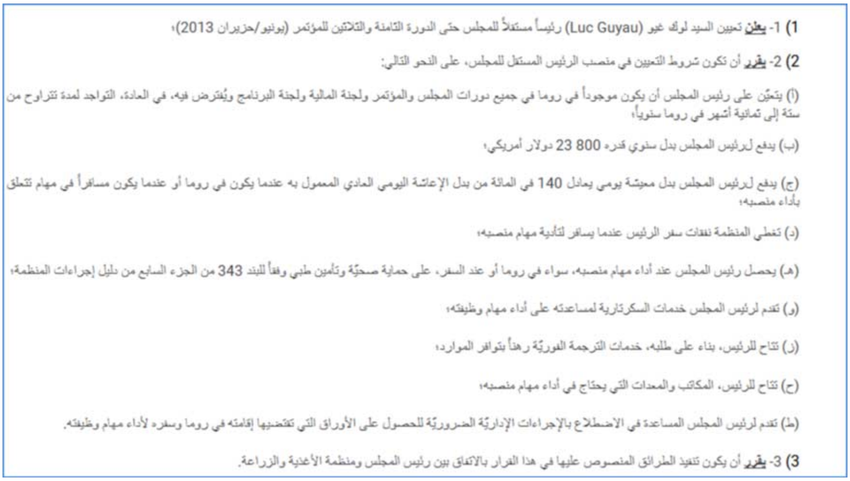
Chinese and Russian variants use Latin letters to itemize list entries, and the problem of synchronization does not arise.

10. Visualizing Normative References

The Akoma Ntoso XML standard makes it easy to produce a network for analysing cross-references across all collections of resolutions. For this reason each resolution is enriched with a graph that shows the inbound and outbound links. Figure 9 shows how Resolution 9/2009 is connected with: blue points are references that cite the current resolution, yellow points are citation included in the text of the current resolution (Figure 10).

11. Conclusions

AKN4UN defined a general framework and guidelines for applying the Akoma Ntoso XML standard to UN documents. In particular a deep analysis was carried out to model the resolutions of UN agencies. FAO was the first UN agency that produced a proof of



1) 1- يعلن تعيين السيد لوك غيو (Luc Guyau) رئيساً مستقلاً للمجلس حتى الدورة الثامنة والثلاثين للمؤتمر (يونيو/حزيران 2013)؛

2) 2- يقرر أن تكون شروط التعيين في منصب الرئيس المستقل للمجلس، على النحو التالي:

(أ) يتعيّن على رئيس المجلس أن يكون موجوداً في روما في جميع دورات المجلس والمؤتمر ولجنة المالية ولجنة البرنامج ويُفترض فيه، في العادة، التواجد لمدة تتراوح من ستة إلى ثمانية أشهر في روما سنوياً؛

(ب) يدفع لرئيس المجلس بدل سنوي قدره 23 800 دولار أمريكي؛

(ج) يدفع لرئيس المجلس بدل معيشة يومي يعادل 140 في المائة من بدل الإعاشة اليومي العادي المعمول به عندما يكون في روما أو عندما يكون مسافراً في مهام تتعلق بأداء منصبه؛

(د) تغطي المنظمة نفقات سفر الرئيس عندما يسافر لتأدية مهام منصبه؛

(هـ) يحصل رئيس المجلس عند أداء مهام منصبه، سواء في روما أو عند السفر، على حماية صحيّة وتأمين طبي وفقاً للبند 343 من الجزء السابع من دليل إجراءات المنظمة؛

(و) تقدم لرئيس المجلس خدمات السكرتارية لمساعدته على أداء مهام وظيفته؛

(ز) تتاح للرئيس، بناء على طلبه، خدمات الترجمة الفوريّة رهناً بتوافر الموارد؛

(ح) تتاح للرئيس، المكاتب والمعدات التي يحتاج في أداء مهام منصبه؛

(ط) تقدم لرئيس المجلس المساعدة في الاضطلاع بالإجراءات الإداريّة الضروريّة للحصول على الأوراق التي تقتضيها إقامته في روما وسفره لأداء مهام وظيفته.

3) 3- يقرر أن يكون تنفيذ الطرائق المنصوص عليها في هذا القرار بالاتفاق بين رئيس المجلس ومنظمة الأغذية والزراعة.

Figure 9. Resolution 13/2011, with a list of points marked using Arabic letters

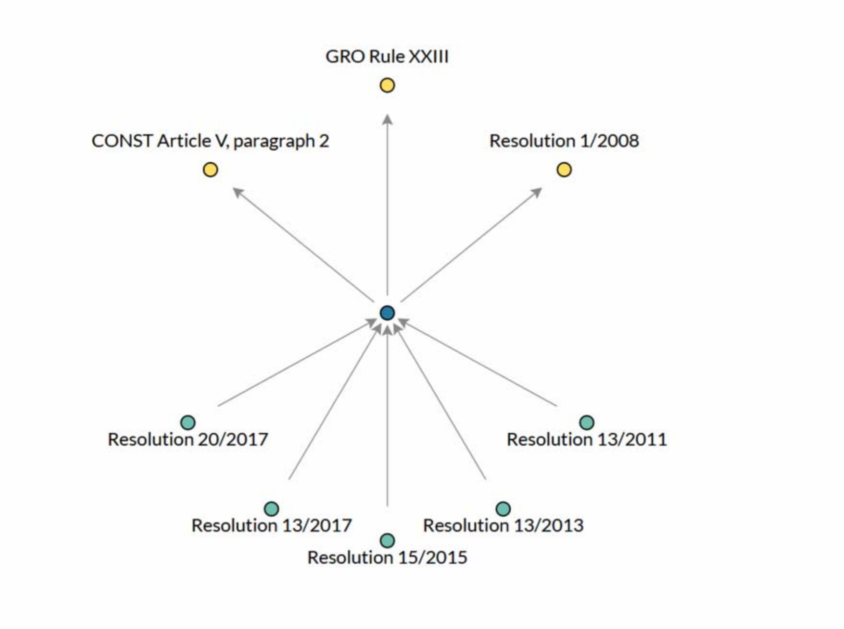


Figure 10. Diagram showing references made to and from Resolution 9/2009

concept is closer to a professional product and ready to go into production. The results of this proof of concept show that the Akoma Ntoso standard is robust in noncanonical legislative domains as well. The irregular structure of some resolutions put Akoma Ntoso under stress, and so we worked out special solutions like using the <crossHeading> element to isolate an odd preambular sentence embedded in the body of the provision. The best enrichment was done using ontologies and Semantic Web annotations to enhance the ability to express the knowledge included in resolutions and to permit specific, effective searching. Also, thanks to the Akoma Ntoso naming convention for IDs and IRIs, the system proved capable of managing multilingualism and dynamic cross-reference navigation between versions of documents at different times and between variants in different languages.

Acknowledgements

This project was cofunded by FAO in the period from 2014 to 2018. We would like to thank the entire FAO team that made it possible to achieve the inspiring results just reported. A special thanks therefore goes to Sergio Ferraro, Jorge Estevez, Gianluca Giulini, Hasmik Ghazaryan, Giona Fumasoli, Tamara Izzo and Maria Grazia Bovo: their contribution was invaluable, and, coupled with their mastery of the domain and their linguistic expertise, it enabled this project to be successfully be brought to completion in a manner consistent with the digital humanities tradition – a standard we would not have been able to meet had the attitude been purely technical.

References

[1] Barabucci, G., Di Iorio, A., Poggi, F. & Vitali, F. (2013). Integration of Legal Datasets: From Meta-model to Implementation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. ACM, 585-594.

[2] Peroni, S., Palmirani, M. & Vitali, F. (2017). UNDO: The United Nations System Document Ontology. In *Proceedings of 16th International Semantic Web Conference*. Springer, 175-183.

[3] Palmirani, M. & Vitali, F. (2011). Akoma-Ntoso for Legal Documents. In Sartor, G., Palmirani, M., Francesconi, E. & Biasiotti, M. A. (Eds.). *Legislative XML for the Semantic Web*. Springer, 75-100.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190019

Language Resources as Linked Data for the Legal Domain

Patricia MARTÍN-CHOZAS,
Elena MONTIEL-PONSODA and Víctor RODRÍGUEZ-DONCEL
Ontology Engineering Group, Universidad Politécnica de Madrid (Spain)

Abstract. This Chapter describes a four-stage methodology to generate Linguistic Linked Data for the legal domain: identification, creation, transformation (to RDF) and linking. The goal of this process is to enhance the presence of legal language resources in the Linguistic Linked Open Data cloud. Since this Chapter is framed within the H2020 LYNX project, aimed at creating a Legal Knowledge Graph, a parallel objective is to employ the resources generated as a linguistic foundation to annotate, classify and translate the legal resources represented in this graph.

Keywords. legal language resources, terminology extraction, linked data, legal knowledge graph

1. Introduction

Originally, *language resources* have been considered as works that collect any type of linguistic information. For the purposes of this work, we define language resources as *pieces of data containing linguistic information in machine readable forms*. There are several types of language resources depending on their format and the type of information represented: glossaries and terminologies (specialised terms), lexical databases (linguistic knowledge for computers), dictionaries (general terms), thesauri (hierarchical controlled vocabularies), etc. Many general dictionaries are available online, such as Merriam Webster[1] and Oxford Dictionary[2]; other terminological resources containing specialised knowledge can also be found on the Internet, such as TermSciences[3] and UNterm[4]. However, language resources for the legal domain are not that present in the Web, since they tend to be owned by legal publishers, thus, not accessible and sometimes published in obsolete and proprietary formats. Moreover, *legal jargon* is intricate and the meaning of terms varies as the legal framework changes. Updating non-machine readable legal glossaries is a time-consuming and difficult task to accomplish. On the other hand, a good understanding of legal terminology is essential to comprehend legal documentation, which also tends to be outdated.

To soften the mentioned hindrances regarding legal terminology and legal documentation, the LYNX project aims at creating a *Legal Knowledge Graph (LKG)*, that is in-

[1] https://www.merriam-webster.com/.
[2] https://www.oxforddictionaries.com/.
[3] http://www.termsciences.fr/termsciences/?lang=en.
[4] https://unterm.un.org/UNTERM/portal/welcome.

terlinking public and private legal resources, metadata, standards and general open data from the legal domain. The idea is to offer access to updated multilingual and multi-jurisdictional legal information. For that purpose, a steady open-access legal language foundation is required.

Linked Data [1] is a particularly convenient form to create such a language cloud, since it is intended to publish interlinked machine-readable data in open-source and non-proprietary formats. In fact, the *Linguistic Linked Open Data (LLOD)*[5] cloud gathers language resources published according to the Linked Data Principles [2], following the RDF set of W3C specifications. Again, within this cloud, legal knowledge is underrepresented. The objective of this contribution is to create the *Linguistic Legal Linked Open Data* (LLLOD) cloud that starts covering the legal knowledge gap. This new legal cloud will set the language foundations to annotate, classify and translate the legal documents taking part of the LKG.

2. Related Work

This Section is divided into three parts: a survey of language resources published as Linked Data, an analysis of those that belong to the legal domain and a summary of the most common models to represent linguistic information.

2.1. Language Resources in RDF

The two main efforts to store and share linguistic information are WordNet [3], a lexical database in English, and BabelNet, an extensive multilingual encyclopedic knowledge base [4]. However, apart from these two knowledge bases, there are many other projects devoted to publishing language resources as per the Semantic Web standards.

One of the most relevant works in this field is the conversion of *IATE* into RDF, the terminological database of the European Union, originally built in TBX (TermBase eXchange format), but unofficially ported to RDF. In this transformation, the *SKOS* vocabulary was used to model term entries with the `skos:concept` property and term relations with `skos:broader` and `skos:narrower` properties. Likewise, *Ontolex* model was used to represent lexical information and term translations thanks to the `ontolex:reference` property [5].

Terminoteca RDF gathers two sets of resources: *Terminesp*, a multilingual terminological database developed by the Spanish Association for Terminology[6]; and terminological glossaries from the *Terminología Oberta* service of the Catalan Terminological Centre[7] (*TERMCAT*). Since the datasets contained in Terminoteca RDF are multilingual, the `vartrans` module of *Ontolex* was a substantial part of the data modeling stage, resulting in a multilingual repository of linked terminologies[8] [6]; [7].

A similar work was made for the conversion into RDF of the bilingual dictionaries used in the *Apertium*[9], free-open-source machine translation system, supported by the Spanish Government and several Spanish universities [8]. The dictionaries are now published as Linked Data following the *lemon* model as part of the *LLOD* cloud [9].

[5] http://linguistic-lod.org/llod-cloud.

[6] http://www.aeter.org/.

[7] http://www.termcat.cat/en.

[8] http://linguistic.linkeddata.es/terminoteca/.

[9] https://www.apertium.org.

Additionally, one of the biggest resources shaping the Linguistic Linked Open Data cloud is *AGROVOC*, a multilingual thesaurus modeled with SKOS-XL and composed by more than 35,000 terms in 29 languages [10]. It has been aligned with other thesauri in the *LLOD*, such as *GEMET* for environment, *TheSoz* for social sciences and *STW* for economics by using the `skos:exactMatch` proprety.

Many other projects are focused on publishing linguistic Linked Data, but these are some of the most representative for this work as they present similar content, models, linked resources and languages.

2.2. Legal Language Resources in RDF

Although the presence of legal language resources in the Web of Data is scarce, some projects have been devoted to improve their representation in the cloud.

One of these projects resulted in the generation of a termbank of multilingual and multi-jurisdictional legal data by linking relevant datasets in the domain such as *IATE*, *Creative Common licenses*, *World Intellectual Property Organization (WIPO) documents*, *DBpedia*[10] and *Lexvo*[11] [11].

Another apposite project in this area is *Eunomos*, a legal knowledge management system based on legislative XML and ontologies [12]. For the extraction and modeling of legal concepts, the system relies on the *Legal Taxonomy Syllabus* ontology, for terminology management of the European Directives [13].

One of the most significant resources taking part in this contribution, *EuroVoc*, has also been represented as Linked Data, following the SKOS vocabulary [14]. This multilingual and multidisciplinary thesaurus created and maintained by the Publications Office of the European Commission is now linked with other sound resources at European level, such as the *UNESCO* and the *GEMET* thesauri. EuroVoc is also available through a *SPARQL endpoint*[12], developed by PoolParty[13].

The Publications Office has also developed the *CELLAR repository* to publish part of the bibliographic resources of the European Union gathered in the *EUR-Lex portal*[14]. Such resources are semantically described by an ontology providing open access, long term preservation, indexing and retrieval services [15]. They continue enhancing semantic interoperability by linking multilingual terminologies to build a *Public Multilingual Knowledge Management Infrastructure* within the *ISA2* project [16].

2.3. Models to Represent Linguistic Information

Some of the resources referenced in Section 2.2 and those that are part of the Linguistic Linked Open Data cloud have been represented following different models, depending on the nature of each resource (structure, content, objectives, etc.). Some of the most common models to represent linguistic information are briefly listed as follows:

- *Lemon*, the *Lexicon Model for Ontologies*, is intended to represent lexical information of a given term, such as the sense, form, abbreviation, etc.[17]

[10] https://wiki.dbpedia.org/.
[11] http://www.lexvo.org/.
[12] https://lynx.poolparty.biz/PoolParty/sparql/Eurovoc4.3.
[13] https://www.poolparty.biz/.
[14] https://eur-lex.europa.eu/.

- *Ontolex* is the evolution of *lemon*. Supported by the W3C Ontology-Lexica Community Group[15], it allows the representation of relations amongst senses, forms and translations [5].
- *LIR*, the *Linguistic Information Repository*, was intended for ontology localisation, offering access to multilingual data [18].
- *Lexinfo* associates additional linguistic information to elements in an ontology [19].
- *SKOS*, the *Simple Knowledge Organization System*, structures thesauri and taxonomies, easing the creation of hierarchical relations between terms. It is widely used within the Semantic Web since it can be combined with formal representation languages, such as the *Web Ontology Language* (OWL) [20].

Choosing the most appropriate vocabulary is decisive for the representation of resources on the Web of Data. However, this task is only one of the steps of a reliable *methodology* that should be followed in order to *publish sound resources as per the Linked Data paradigm*. Such methodology stresses the importance of preprocessing the data, choosing a sound URI naming strategy, selecting the right technology for RDF generation and reliably linking with other datasets in the cloud [21].

3. Motivation

The aforementioned methodology also emphasizes the importance of reusing resources. When searching for datasets to reuse in this work, *a lack of legal language resources* has been identified, specifically in the three subdomains that are of our interest in LYNX project: labour law, data protection and industrial standards. To help represent legal domain in the Web of Data, *new linguistic resources need to be created* from scratch by extracting terms from legal corpora.

While the advantages of publishing Linked Data are increasingly gaining attention, the greatest part of the available language resources nowadays are *not in machine readable formats yet*. Furthermore, although governments and public institutions are publishing legislation on the Internet, they barely apply open format standards and generate the most part of the documents in PDF. This practice causes huge hindrances when updating, sharing and reusing resources. For this reason, it is required to *transform these resources into RDF* so they can be interlinked to provide a more efficient access to multilingual and multi-jurisdictional legal information.

It is worth mentioning that many of the linguistic portals and repositories consulted throughout this work were not updated or maintained any longer. *Documentation of resources* is key to store and share up-to-date knowledge. Part of this contribution has also focused on the creation of a data portal to keep track of all resources handled in this work and specially those shaping the first approach of the *LLOD* cloud.

4. Contribution

The four stages in which this methodology is divided cover the gaps mentioned in Section 3 and are structured as follows:

- *Identification of existing language resources* that could be reused in this project.

[15] https://www.w3.org/community/ontolex/.

- *Generation of new of language resources* by the extraction of terms from legal corpora provided by LYNX partners.
- *Conversion into RDF* of several language resources identified in the first stage and others created afterwards.
- *Linking the resulting terminological resources* from the previous stages with other existing datasets in the LLOD.

4.1. Identification of Existing Resources

Three different search strategies have been explored with the aim of spotting potentially useful resources:

- Identification through general web search
- Identification through literature
- Identification through specialised portals (ELRC-SHARE[16], Retele[17], CLARIN[18] and the OLAC Language Resource Catalog[19], amongst others).

Besides AGROVOC, IATE, EuroVoc, STW or GEMET, mentioned in Section 2, several significant language resources have also been identified. The *German Labour Law Thesaurus*, for instance, covers different areas of labour law and it is published as Linked Data. *JuriVoc* is a multilingual thesaurus containing juridical terms hierarchically structured. Likewise, the *TERMCAT* institute published glossaries from the labour law domain in XML, a very convenient format to be transformed into RDF, since they present similar structures.

Therefore, Table 1 gathers the first set of available language resources that could be reused in this project. It is worth mentioning that they present different formats and that not all of them contain legal information but cover many adjacent domains. In addition, some resources from the general domain have also been gathered since they present other interesting features (updates, links with other datasets, etc.).

4.2. Creation of New Resources

The approach followed here can be divided into three different stages:

- Evaluation of term extraction tools.
- Extraction of terms from legal corpora.
- Evaluation of extracted terms.

For the creation of resources, the first task consisted in identifying available *Automatic Term Extraction (ATE)* tools to be tested, in order to choose the one that met the needs of this work. Nine ATE tools have been evaluated: Translated.net[20], VocabGrabber[21], TermSuite[22], TermoStat Web[23], SketchEngine[24], Fivefilters[25], Termine[26], Pootle[27]

[16] https://www.elrc-share.eu/.

[17] http://catalogo.retele.linkeddata.es/.

[18] https://www.clarin.eu/.

[19] http://www.language-archives.org/.

[20] https://labs.translated.net/terminology-extraction/.

[21] https://www.visualthesaurus.com/vocabgrabber/.

[22] http://termsuite.github.io/.

[23] http://termostat.ling.umontreal.ca.

[24] https://www.sketchengine.co.uk/.

[25] http://fivefilters.org/term-extraction/.

[26] http://www.nactem.ac.uk/software/termine.

[27] https://pootle.translatehouse.org/.

and TBXTools[28]. Eight evaluation criteria were considered, including availability of the tool, file formats, type of extracted terms and additional services. After several extraction tests per tool and based on the quality of the results as assessed by expert terminologists, *SketchEngine* was the tool selected to generate new glossaries.

Table 1. Set of available language resources identified

ID	Name	Description	Language
iate	IATE	EU terminological database	EU languages
eurovoc	Eurovoc	EU multilingual thesaurus	EU languages
eur-lex	EUR-Lex	EU legal corpora portal	EU languages
conneticut-legal-glossary	Connecticut Legal Glossary	Bilingual Legal Glossary	en, es
unesco-thesaurus	UNESCO Thesaurus	Multilingual multidisciplinary thesaurus	en, es, fr, ru
library-of-congress	Library of Congress	Legal corpora portal	en
imf	International Monetary Fund	Economic multilingual terminology	en, de, es
eugo-glossary	EUGO Glossary	Business monolingual dictionary	es
cdisc-glossary	CDISC Glossary	Clinical monolingual	en
stw	STW Thesaurus for Economics	Economic monolingual thesaurus	en
edp	European Data Portal	EU datasets	EU languages
inspire	INSPIRE Glossary (EU)	General terms and definitions in English	en
saij	SAIJ Thesaurus	Controlled list of legal terms	es
calathe	CaLaThe	Cadastral vocabulary	en
gemet	GEMET	General multilingual thesauri	en, de, es, it
informea	InforMEA Glossary (UNESCO)	Monolingual glossary on environmental law	en
copyright-termbank	Copyright Termbank	Multi-lingual term bank of copyright-related terms	en, es, fr, pt
gllt	German labour law thesaurus	Thesaurus with labour law terms	de
jurivoc	Jurivoc	Juridical terms from Switzerland	de, it, fr
TERMCAT	TERMCAT	Terms from several fields including law	ca, en, es, de, fr, it
termcoord	Termcoord	Glossaries from EU institutions and bodies	EU languages
agrovoc	Agrovoc	Controlled general vocabulary	29 languages

Terms were extracted from three different corpora provided by LYNX partners, each one representing one area of law: a set of collective agreements from the *labour law domain*, a set of regulations from the *data protection* domain and a set of decisions from the *industrial standards* domain.

As a result, the tool returned a list of 200 candidate terms, single and multi-word, per set of documents. The lists were evaluated by professional terminologists to analyse the quality of the result. Such evaluation has been performed by verifying the terms against well-known terminological databases widely used by language professionals: IATE[29], Linguee[30] and also BabelNet[31].

These checks showed that several candidates were not correctly identified by the tool and part of those that were correctly extracted are not relevant for the legal domain. Therefore, Table 2 gathers the amount of ‘clean terms’ after the evaluation of each new glossary.

The glossaries have been organised in XLS files with eight columns per entry. Each column represents an attribute of the term (URI, definition, usage note, etc.) that will be represented in the conversion stage as an RDF property.

[28] https://sourceforge.net/projects/tbxtools/.

[29] https://iate.europa.eu/home.

[30] https://www.linguee.es/.

[31] https://babelnet.org/.

Table 2. Term lists after the evaluation stage

Resulting term lists		
Labour Law Glossary (ES)	Data Protection Glossary (EN)	Industrial Standards Glossary (EN)
102 terms	98 terms	109 terms

4.3. Conversion into RDF

The objective of this stage is to transform into RDF the glossaries created in the previous Section and some of the resources identified in Section 4.1. As stated in the mentioned Section, TERMCAT platform gathers glossaries in XML from the labour law domain. Since the LYNX project handles data from this domain, two of them have been selected to be reused, converted and linked in this work. Consequently, five glossaries have been converted into RDF: three new resources created by term extraction from legal corpora (two in English and one in Spanish) and two existing TERMCAT glossaries (one in English and one in Spanish).

Due to the nature, goal and format of the glossaries, SKOS was the vocabulary selected to transform the resources, since it is an intuitive model able to represent all the term attributes contained in the glossaries.

Therefore, Figure 1 exemplifies the representation of a term entry and its attributes as they are structured in the glossaries.

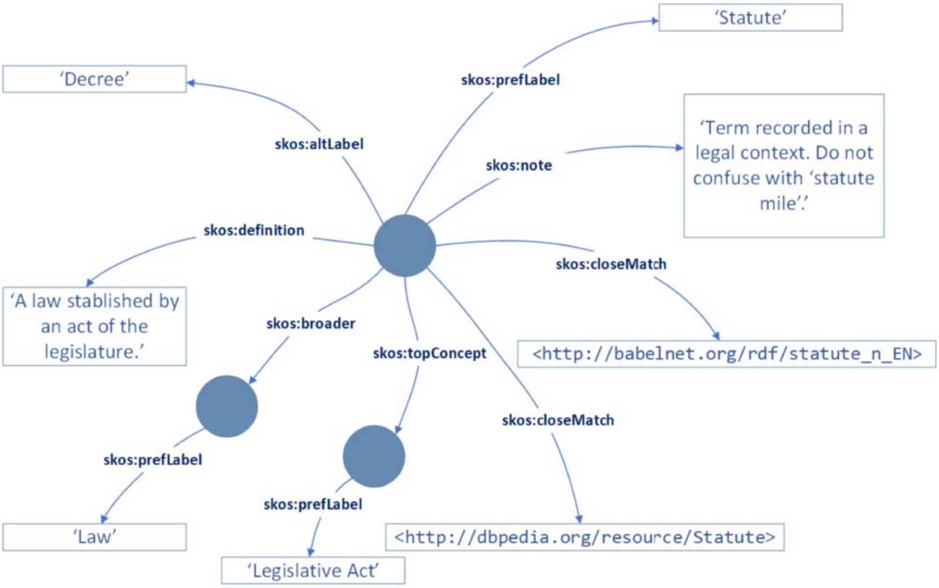

'Statute'
'Decree'
skos:prefLabel
skos:altLabel
'Term recorded in a legal context. Do not confuse with 'statute mile'.'
skos:note
skos:definition
skos:closeMatch
'A law stablished by an act of the legislature.'
skos:broader
skos:topConcept
<http://babelnet.org/rdf/statute_n_EN>
skos:closeMatch
skos:prefLabel
skos:prefLabel
'Law'
<http://dbpedia.org/resource/Statute>
'Legislative Act'

Figure 1. Representation of a term entry with SKOS

The URI of the term is represented by the `skos:concept`. It has been generated by following the URI naming strategy of related work in RDF generation such as the conversion of TERMCAT files [6] and the Apertium Bilingual Dictionaries [8]. These approaches use URL of the server were the resource is located, the part of speech of each term and the ISO 639-1 language code to build each identifier. An example of a term entry in the glossaries handled here is the following:

http://linguistic.linkeddata.es/terminoteca/lynx/statute-n-en

Finally, the metadata of each glossary (author, creation date, title and description of the resource, etc.) were modeled with the DublinCore[32] ontology.

4.4. Linking Step

Once the glossaries have been represented in RDF, the next step is to generate hyperlinks that connect them with other knowledge bases and linguistic linked resources that are already part of the Linguistic Linked Open Data cloud. These links between resources are exceptionally helpful to share information and enrich the glossaries with context, translations, related terms, usage notes, etc.

Both conversion and linking processes have been performed with OpenRefine[33]. The linking service of OpenRefine can be executed by using either a SPARQL endpoint or an RDF dump of the knowledge base. In this case, the SPARQL endpoint option has been applied since the three involved knowledge bases offer this kind of access: *DBpedia*[34], *EuroVoc*[35] and *BabelNet*[36].

Table 3 shows the results of the linking experiments:

Table 3. Results of the linking tests

Glossaries	BabelNet		EuroVoc		DBpedia		Total terms
Labour Law Glossary ES	47	46%	-	-	8	7.84%	102
Data Protection Glossary EN	70	71%	13	13.27%	59	60.20%	98
Industrial Standards EN	37	33.94%	-	-	70	64.22%	109
Termcat Glossary ES	-	-	97	13.18%	61	8.29%	736
Termcat Glossary EN	-	-	104	13.90%	118	15.78%	748

These results show that the number of links generated is highly dependent on the content of the glossaries and on the type of knowledge base to which they have been linked with. For instance, the lowest percentages appear when linking labour law glossaries with DBpedia, while more technical domains are well represented.

Figure 2 represents the result of this work as the first approach of the Linguistic Legal Linked Open Data cloud (LLLOD).

All the language resources appearing in Figure 2 are duly documented in LYNX Data Portal[37] and all of them can be openly accessed. The number of datasets of the portal is constantly increasing as the LYNX project progresses.

[32] http://dublincore.org/.

[33] http://openrefine.org/.

[34] https://wiki.dbpedia.org/.

[35] http://eurovoc.europa.eu/.

[36] https://babelnet.org/.

[37] http://data.lynx-project.eu/.

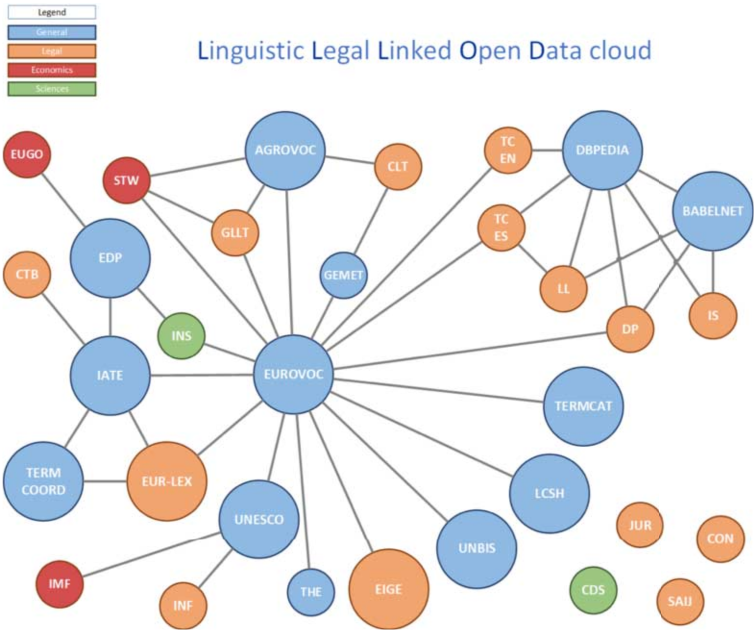


Figure 2. First approach of the Linguistic Legal Linked Open Data cloud

5. Conclusions and Future Work

Legal domain undergoes a lack of language resources in structured and open formats. From the whole set of resources identified, the 39% corresponds to archived resources that might be relevant for the domain but published in obsolete formats or not supported any longer. The next steps on this matter will be focused on publishing the resources identified in the first stage of this methodology as Linked Data to enrich the LLOD.

During the creation of new resources, it was noticed that performance of ATE tools for the legal domain is still limited. These experiments showed that a 40% of the candidate terms were not correctly identified, which means that a huge amount of manual work is still required, unaffordable in large projects. More research on this field is also required in order improve the accuracy of ATE tools on legal corpora and the automation of the whole extraction process.

On the other hand, the current glossaries generated contain monolingual information represented with SKOS vocabulary. In order to add value to these resources, term translations and additional information will also be included. Therefore, other RDF models need to be considered for the representation of this type of relations (e.g. Ontolex `vartrans` module) [22].

Likewise, as mentioned in Section 4, linking tools present several drawbacks that also involve a huge amount of manual work. Consequently, other tools for publishing Linked Data, such as VocBench[38] and Silk[39], will also be tested.

[38] http://vocbench.uniroma2.it/.
[39] http://silkframework.org/.

Acknowledgements

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the contract LYNX (grant agreement no.: 780602).

References

[1] Berners-Lee, T. (2006). *Design Issues*, https://www.w3.org/DesignIssues/LinkedData.html.

[2] Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems, 5*(3), 1-22.

[3] Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39-41.

[4] Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence, 193*, 217-250.

[5] Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V. et al. (2015). Linked Terminologies: Applying Linked Data Principles to Terminological Resources. In *Proceedings of the eLex 2015 Conference*, 504-517.

[6] Bosque-Gil, J., Montiel-Ponsoda, E., Jorge, G. & Aguado-de-Cea, G. (2016). Terminoteca RDF: A Gathering Point for Multilingual Terminologies in Spain. In Erdman Thomsen, H., Pareja-Lora, A. & Nistrup Madsen, B. (Eds.). *Term Bases and Linguistic Linked Open Data. Proceedings of the 12th International Conference on Terminology and Knowledge Engineering*, 136.

[7] Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. (2016). Linked Data in Lexicography. *Kernerman Dictionary News, 24*, 19-24.

[8] Gracia, J., Villegas, M., Gómez-Pérez, A. & Bel, N. (2017). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web, 9*(2), 231-240.

[9] McCrae, J., Spohr, D. & Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In Antoniou, G., Grobelnik, M., Simperl et al. (Eds.). *The Semantic Web: Research and Applications. Proceedings of the 8th Extended Semantic Web Conference*. Springer, 245-259.

[10] Caracciolo, C., Stellato, A., Morshed, A. et al. (2013). The AGROVOC Linked Dataset. *Semantic Web, 4*(3), 341-348.

[11] Rodríguez-Doncel, V., Santos, C., Casanovas, P. & Gómez-Pérez, A. (2015). A Linked Term Bank of Copyright-related Terms. In Rotolo, A. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 28th Jurix Conference*. IOS Press, 91-100.

[12] Boella, G., Humphreys, L., Martin, M., Rossi, P., van der Torre, L. & Violato, A. (2012). Eunomos, a Legal Document and Knowledge Management System for Regulatory Compliance. In De Marco, M., Te'eni, D., Albano, V. & Za, S. (Eds.). *Information Systems: Crossroads for Organization, Management, Accounting and Engineering*. Springer, 571-578.

[13] Ajani, G., Lesmo, L., Boella, G., Mazzei, A. & Rossi, P. (2007). Terminological and Ontological Analysis of European Directives: Multilinguism in Law. In Gardner, A. & Winkels, R. (Eds.). *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. ACM, 43-48.

[14] Díez, L. A., Pérez-León, B., Martínez-González, M. & Blanco, D. J. (2010). Propuesta de representación del tesauro Eurovoc en SKOS para su integración en sistemas de información jurídica. *Scire: representación y organización del conocimiento, 16*(2), 47-51.

[15] Francesconi, E., Küster, M. W., Gratz, P. & Thelen, S. (2015). The Ontology-based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. In Kö, A. & Francesconi, E. (Eds.). *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 29-39.

[16] Schmitz P., Francesconi E., Hajlaoui N. & Batouche B. (2017). Towards a Public Multilingual Knowledge Management Infrastructure for the European Digital Single Market. In *1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets, Proceedings of the LDK 2017 Workshops*, Vol-1899, p. 33-42.

[17] McCrae, J., Aguado-de-Cea, G., Buitelaar, P. et al. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation, 46*(4), 701-719.

[18] Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A. & Peters, W. (2011). Enriching Ontologies with Multilingual Information. *Natural Language Engineering, 17*(3), 283-309.

[19] Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web, 9*(1), 29-51.

[20] Miles, A. & Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System Reference*. W3C recommendation, 18.

[21] Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. & Aguado-de-Cea, G. (2014). Publishing Linked Data on the Web: The Multilingual Dimension. In Buitelaar, P. & Cimiano, P. (Eds.). *Towards the Multilingual Semantic Web*. Springer, 101-117.

[22] Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. & Montiel-Ponsoda, E. (2015). Applying the Ontolex Model to a Multilingual Terminological Resource. In Gandon, F., Guéret, C. & Villata, S. et al. (Eds.). *The Semantic Web: ESWC 2015 Satellite Events. Proceedings of the European Semantic Web Conference*. Springer, 283-294.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190020

DaPIS: An Ontology-Based Data Protection Icon Set

Arianna ROSSI [a,b,1] and Monica PALMIRANI [b]
[a] *SnT, University of Luxembourg*
[b] *CIRSFID, Università degli Studi di Bologna (Italy)*

Abstract. Privacy policies are known to be impenetrable and lengthy texts that are hardly read and poorly understood. This is why the General Data Protection Regulation (GDPR) introduces provisions to enhance information transparency including icons as visual means to clarify data practices. However, the research on the creation and evaluation of graphical symbols for the communication of legal concepts, which are generally abstract and unfamiliar to laypeople, is still in its infancy. Moreover, detailed visual representations can support users' comprehension of the underlying concepts, but at the expense of simplicity and usability. This Chapter describes a methodology for the creation and evaluation of DaPIS, a machine-readable Data Protection Icon Set that was designed following human-centered methods drawn from the emerging discipline of Legal Design. Participatory design methods have ensured that the perspectives of legal experts, designers and other relevant stakeholders are combined in a fruitful dialogue, while user studies have empirically determined strengths and weaknesses of the icon set as communicative means for the legal sphere. Inputs from other disciplines were also fundamental: canonical principles drawn from aesthetics, ergonomics and semiotics were included in the methodology. Moreover, DaPIS is modelled on PrOnto, an ontology of the GDPR, thus offering a comprehensive solution for the Semantic Web. In combination with the description of a privacy policy in the legal standard XML Akoma Ntoso, such an approach makes the icons machine-readable and automatically retrievable. Icons can thus serve as information markers in lengthy privacy statements and support an efficient navigation of the document. In this way, different representations of legal information can be mapped and connected to enhance its comprehensibility: the lawyer-readable, the machine-readable, and the human-readable layers.

Keywords. data protection, icons, semantic technologies, ontology, transparency, legal design

1. Introduction

Traditionally, the protection of data subjects in the EU has employed the fundamental tool of mandated disclosure about the collection, use and sharing of their personal data. By be informed about the existence of such data practices, individuals would be in control of their personal information and would be able to make an informed choice either to use or not a certain service (i.e. informed consent) [1]. However, research (e.g., [1]; [2]; [3];

[1] This research was carried out while Arianna Rossi was a fellow at the ICR, University of Luxembourg and was supported by the LAST-JD PhD program (http://www.last-jd.eu/) financed by the EACEA.

[4]; [5]) and empirical evidence have revealed the poor implementation of the principle of transparency of information: often privacy policies serve the mere function of legal compliance, rather than fulfilling their supposed informative function. As much other legal communication [6], "[p]rivacy policies are written by lawyers, for lawyers, and appear to serve little useful purpose for the data subject due to their length, complexity and extensive use of legal terminology" ([7], p. 29). As a consequence, privacy policies are hardly read and insufficiently understood by data subjects.

Although some scholars support the complete abandon of mandated disclosures [8], the General Data Protection Regulation (GDPR)[2] has challenged this assumption: it has imposed transparency of information as a legal obligation on the data controller and has provided very specific indications about how to translate such abstract requirement into practical and implementable solutions. The quality of the information assumes an unprecedented importance to demonstrate compliance: plain language must be employed in any communication addressed to data subjects (Article 12.1) to overcome the use of overly complex language in lengthy texts that discourage individuals from reading (see e.g. [9]). Moreover, Article 12.7 introduces the possibility to accompany the information provided to data subjects with standardized icons to present "in an easily visible, intelligible and clearly legible manner a meaningful overview of the intended processing". Such icons must be machine-readable if employed in electronic format.

Although eventually it will be the role of the European Commission to adopt delegated acts to guide the creation of these icons, the need of expert advice is emphasized in Recital 166 GDPR and in the dedicated Guidelines on Transparency by the Article 29 Working Party[3], which encourages an 'evidence-based approach' and 'extensive research' ([10], p. 26) to inform the development and application of the icons, and to determine their efficacy. The investigation presented in this Chapter intends to provide a contribution to the (still) scarce academic research about such topic and to provide the foundations for further investigation.

2. Research Scenario

The GDPR's call for machine-readable graphical elements to express data practices identifies two relevant intertwined lines of research. On the one hand, there are the technologies for the management and (semi-)automated extraction of information from legal documents that provide information interpretable by machines (explored in Section 3). On the other hand, there are interventions that aim to facilitate humans' accessibility to legal information to tackle the problems outlined above: namely, a specific area of Legal Design, i.e. the human-centered design of legal information (illustrated in Section 4). Without an interface, machine-readable information is confined to the exclusive world of computers and technical experts, whereas user-friendly and visualized documents are not meaningful for machines [11]. DaPIS, the Data Protection Icon Set modeled on a computational ontology of the GDPR, is an attempt to reconcile these two directions of research. This Chapter details a methodology to create a 'visual layer' from marked-up privacy

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1, 4.5.2016.

[3] Now the European Data Protection Board.

policies that complement the well-established levels of legal information representation [12], with the aim of communicating data practices in an user-friendly manner.

Moreover, the integration with technology also has the goal to reproduce at scale and efficiently those good practices of Legal Design that have proved useful and valuable in specific contexts. Many human-centered solutions applied to legal content are unique and crafted *ad hoc* for specific users in specific contexts (e.g. in contract design, see e.g. [13]; [14]). Instead, information for data subjects and for consumers is mainly addressed to unspecified audiences and the drafters do not usually now the demographics of their potential readers, thus it becomes difficult to design good user-oriented communications. Although there does not always exist a one-fits-all solution, this challenge motivates the investigation and establishment of best practices that can be generalized, or even standardized. Legal Design patterns (i.e. repeatable, systematized and extensible solutions to recurring problems in the legal domain) [15]; [16]; [17]; [18] can represent a viable manner to support the concrete implementation of abstract legal principles [19], like the principle of transparency, as opposed to a jungle of bespoke solutions. The ontology-based companion icons presented in the next paragraphs correspond to one of those patterns [20].

The next Sections illustrate the methodology that was designed to create DaPIS and integrate it with existing semantic technologies for the legal domain (see also [21]). Section 3.1 exemplifies the multi-layered structure that characterizes the machine-interpretable representation of legal documents and adds an additional visual layer. Section 3.2 describes the design of PrOnto, i.e. an ontology of the GDPR, and its constitutive modules, that provided DaPIS' objects of representation. Then, Section 4.1 briefly introduces Legal Design, while in the following Section 4.2 the iterative cycle of icon design is illustrated, alongside the importance of multi-disciplinarity to solve such challenging task (Section 4.3) and the features of modularity and compositionality that constitute the icon set are exposed (Section 4.4). Lastly in Section 4.5, the iterative evaluation phases are shortly illustrated. The final Section 5 provides indications about possible future directions of research.

3. Integrating Machine-Readable and Visual Representations for the Legal Domain

3.1. A Multi-layered Architecture: Adding a Visual Layer

The adoption of semantic technologies in the legal domain envisages a multi-layered architecture for the formal representation and the management, maintenance and communication of the information contained in legal documents. Indeed, different information is expressed through strictly separate layers: text; structure; legal metadata; legal ontology; legal rules [12].

The first three levels can be implemented through the Akoma Ntoso XML schema which is nowadays a well-established international standard adopted by many institutions around the world. Akoma Ntoso enables the addition of descriptive structure (i.e., the structural and semantic mark-up) to the content of such documents [12]. The metadata level (i.e. the third layer) adds descriptions about the content of the legal document [22]: the tags of the inline semantic mark-up can be linked to a reference in the metadata section that, in turn, points to an external resource that defines the meaning of such tags [12], namely an ontology (i.e. the fourth layer). Lastly, the fifth layer provides the legal meaning of the text and transforms the norms into rules to allow, for example, automated reasoning through machine-interpretable languages like LegalRuleML.

Figure 1 illustrates the application of such multi-layered architecture structure to privacy policies. When the text of a privacy policy is marked up with tags linked to instances of a specific ontology, its semantic content can be described in a univocal and machine-interpretable fashion. The concepts of the ontology can be associated to their corresponding icon and, hence, be semi-automatically summoned by the semantic tags[4]. The underlying assumption, derived from [13]; [16], hypothesizes that icons can accompany the text to clearly indicate where a specific information item appears in long privacy policies, thereby supporting the activity of information finding of the reader in a quicker and more effective manner (i.e. 'companion icons').

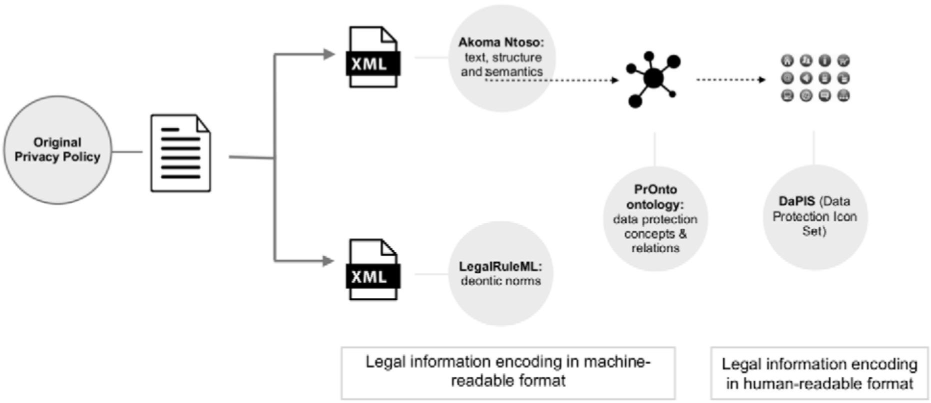

Original Privacy Policy
XML
Akoma Ntoso: text, structure and semantics
XML
LegalRuleML: deontic norms
PrOnto ontology: data protection concepts & relations
DaPIS (Data Protection Icon Set)
Legal information encoding in machine-readable format
Legal information encoding in human-readable format

Figure 1. The multi-layered architecture of a privacy policy [12] combined with an additional visual dimension

The description of legal information in a machine-interpretable format also allows automated reasoning on legal texts, for example to draw inferences and match expressions in natural language to the corresponding ontological instance (e.g. the expressions 'you' and 'user' can both refer to the concept of data subject). Moreover, an ontology is independent from language, which counts as an additional strength: the same icon can be provided for text spans expressed in different languages, which yet refer to the same ontological concepts, whilst correspondent labels in different languages can be provided for the same icon.

For such reasons and following the Regulation's call for 'machine-readable' icons on electronical devices, DaPIS (see Section 4) has been fashioned on the conceptual modules of the computational ontology PrOnto. Figure 2 illustrates the network among the document, its mark-up in Akoma Ntoso, the metadata references pointing at the external ontology and the human-oriented representation characterized by information architecture and companion icons.

[4] Provided the development of such a tool, which was not the goal of the project described in these pages, though.

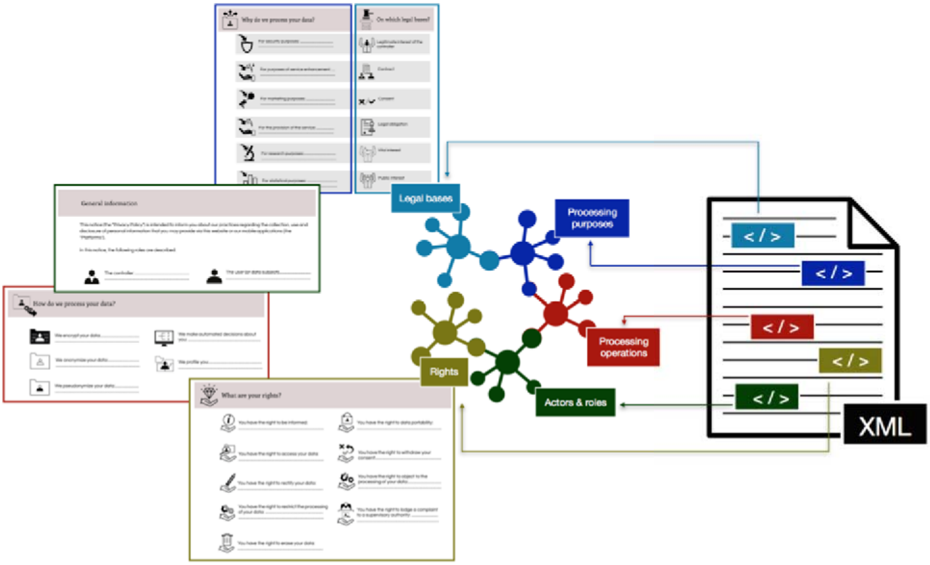


Figure 2. The semantic network among the Akoma Ntoso mark-up of a privacy policy (on the right-hand side), the corresponding concepts arranged in classes of the PrOnto ontology (in the middle) and a visual layer characterized by information architecture and companion icons (on the left-hand side)

3.2. PrOnto

This Section briefly introduces PrOnto[5], the computational ontology that formally models the concepts and norms contained in the GDPR and that has described in detail elsewhere [23]; [24]; [25]. The GDPR provides a European legal framework that defines concepts of data protection, relations among them, and a common vocabulary to describe them. PrOnto is mainly aimed at helping companies and organizations to comply with the many obligations set forth by the GDPR. Combined with other Semantic Web technologies and legal reasoners, goal of this ontology is to facilitate the data controllers' fulfillment of their duties, such as the undertaking of the Data Protection Impact Assessment and the detection of those violations (e.g. a data breach) that envisage countermeasures. Yet, the GDPR constitutes only the initial, central core of norms that have been modeled in Pronto, which is meant to be expanded to other legal frameworks and jurisdictions.

3.2.1. The Design of PrOnto

PrOnto has been designed by following MeLOn [23], an interdisciplinary methodology for the creation of legal ontologies, which is composed of a series of recursive steps. In the first place, the research questions that the ontology aims to address and practical use-cases for its eventual application have been defined: namely, modeling the legal norms defined in the GDPR to allow legal reasoning and compliance checking. Thus, PrOnto has put an emphasis on the modeling of the processing operations, and of the obligations

[5] Developed within the FNR/CORE DAPRECO (DAta Protection REgulation COmpliance https://www.fnr.lu/projects/data-protection-regulation-compliance/) project, at the University of Luxembourg and University of Bologna.

and rights belonging to the different roles (e.g., data subject, controller, etc.) defined by the Regulation.

Then, the GDPR was analyzed by a team of legal knowledge engineers to extract relevant concepts and relations among them, e.g. the different stakeholders affected by this Regulation and their respective rights and duties. This knowledge was then integrated with expert feedback and additional information taken from other authoritative sources, such as Opinions and Guidelines from the Article 29 Working Party (e.g. [26]) and guidance from the UK's Information Commissioner's Office[6], as well as international standards[7]. Moreover, best practices for ontological knowledge modeling have been followed: PrOnto is framed in foundational and core ontologies such as ALLOT [27], FRBR [28], LKIF-core [29], and PWO [30]. In addition, ontology design patterns that express values in time and context [31] have been reused.

MeLOn also provides for the evaluation of the ontology in application to concrete use-cases in terms of coherence, completeness, efficiency, effectiveness, agreement, and usability. Lastly, a testing phase that makes use of the OntoClean method [32] and of SPARQL queries establishes if the research goals defined at the beginning of the ontology design have been reached. PrOnto is currently being employed in projects where its capacity to indicate legal compliance in a variety of scenarios is being assessed [25]. Publication and feedback collection is the last step that contribute to reach a shared agreement within the community of legal experts.

3.2.2. PrOnto's Modules

In the following PrOnto's conceptual modules that constituted the object of DaPIS's design (see Section 4) are described:

1. *Data and documents*: personal data (as opposed to non-personal data and anonymized data), sensitive data, and the documents (e.g. privacy policies, DPIAs, contracts, etc.) that describe and regulate the relationships among different actors involved in the processing;
2. *Agents and roles*: agents can play multiple roles depending on the context and the processing operation (e.g. the same person can be a data subject in one context and a data controller in another one) which also determine their rights and duties;
3. *Processing operations*: these are modelled through a workflow [30], i.e. a sequence of steps with a specific input and a specific output. The essential actions in data processing that were rendered graphically are: anonymize (subclass of delete), pseudonymize (subclass of derive), automated decision-making (individual of infer, specified with a boolean data property), profiling, direct marketing, encrypt, copy, and transfer of personal data to third countries (individual of the class transmit, specified with a place axiom);
4. *Deontic Operators*: the legal norms are modeled in terms of deontic operators (i.e. rights, obligations, permissions, and prohibitions), in order to be integrated with LegalRuleML to support compliance checking with the GDPR. In the perspective of transparency, the rights of the data subject (Articles 12-22) assume paramount importance;

[6] ICO (2014). *Deleting Personal Data*. Technical report.

[7] ISO (2018). *ISO 31000:2018 – Risk management*.

5. (a) *Processing purposes*: the principle of lawfulness (Article 6) establishes that personal data processing must be motivated by specific purposes, that were extracted from articles and recitals of the normative text; (b) *Legal bases*: every purpose must be supported by one of the possible legal bases laid down in Article 6: consent, contract, legal obligation, public interest, vital interest, legitimate interest. Note that the consent and the contract are subclasses of the document class.

Table 1 lists DaPIS' icons in correspondance of the classes and subclasses composing PrOnto.

Table 1. Classes and subclasses of PrOnto that have been visualized in DaPIS

Superclass	Class
(1) Data	Personal Data
(2) Agents' roles	Data subject Data controller Supervisory authority
(3) Processing operations	Copying Pseudonymization Anonymization Direct marketing Automated decision-making Profiling Encryption Transfer of personal data to third countries Storage of personal data in the EU Data sharing with third parties
(4) Data subject's rights	Right to be informed Right of access Right to rectification Right to erasure Right to withdraw consent Right to data portability Right to restriction of processing Right to object to processing Right to lodge a complaint
(5a) Processing purposes	Research purposes Statistical purposes Purpose of information security Purpose of provision of the service Purpose of service enhancement Marketing purposes Profiling purposes
(5b) Legal bases for processing	Consent Legal obligation Vital interest Public interest Legitimate interest Contract

In conclusion, unlike other data protection icon sets focusing on data types and a handful of processing operations (e.g. [33]; [34]), DaPIS provides a systematic formalization and classification of data protection concepts and relations among them. It additionally covers legal bases for processing, rights of the data subject and a more extended set of processing purposes. These information items ought to be presented to data subjects when processing takes place according to Articles 13 and 14 GDPR. Although some of the other icon sets partially overlap with the one described in these pages and can thus be successfully integrated, DaPIS also provides graphical representations for concepts that have never been visualized, like the rights of the data subject.

4. DaPIS: The Data Protection Icon Set

The project described in these pages has used the structured and formalized representation of the data protection domain explained above to define and circumscribe the items meant to be visualized. Legal Design methods were then employed for the creation of the icon set.

4.1. Legal Design and Users of the Law

Legal Design is an interdisciplinary approach for "the application of human-centered design to the world of law, to make legal systems and services more human-centered, usable, and satisfying" ([35], Chap. 1). Human-centered design focuses on the development of solutions that consider the target audience's needs: in this view, 'users of the law' are not only lawyers, judges and regulators, but also citizens, businesses and laypeople in general. For this reason, Legal Design favors interdisciplinarity and participatory design approaches [36], with the aim to recompose the fracture between the theoretical assumptions of the law and actual individual's need and abilities.

In the case of the discrepancy briefly analyzed in Section 1, the law provides for instruments (i.e. mandated disclosures) that should guide the data-related choices of data subjects in an autonomous and informed manner. Yet, the implementation of the principle of transparency is generally so poor and problematic that individuals do not view privacy policies as an empowering instrument to understand their rights and how legal rules apply to them, but rather as a nuisance that is ignored or rapidly clicked away. As they are traditionally fashioned, privacy terms serve the needs of lawyers and regulators, but not those of data subjects. Indeed, a wall of text without any affordance for the human eye discourage individuals from engaging with the reader. On the contrary, visual cues can demonstrably attract reader's attention, reinforce memorization and support effective information finding.

4.2. An Iterative Architecture for the Design of DaPIS

Legal Design includes participatory design methods in its toolkit in order to consider and involve all the users of the law. This is why, DaPIS was created through a series of multi-stakeholders' workshops that followed the design cycle phases: 1. discover; 2. synthesize; 3. build; 4. test; 5. evolve [35]. Three working versions of DaPIS were created and evaluated in an iterative manner: the icons of each version were evaluated in a user study (see Section 4.5) and, if needed, consequently refined in the following workshop(s). Figure 3 exemplifies the DaPIS design cycle.

The development of DaPIS can be described as follows:

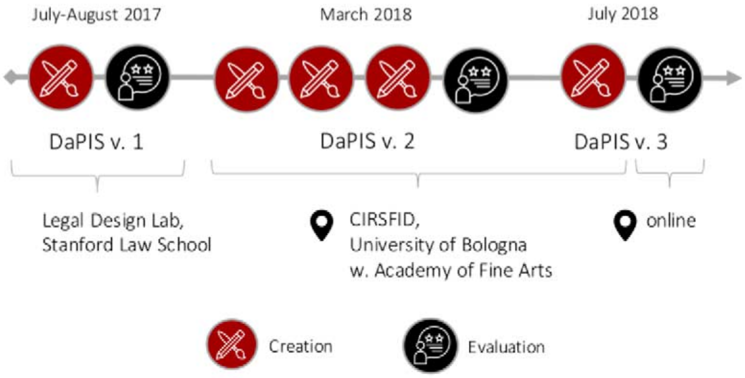

July-August 2017
March 2018
July 2018
DaPIS v. 1
DaPIS v. 2
DaPIS v. 3
Legal Design Lab, Stanford Law School
CIRSFID, University of Bologna w. Academy of Fine Arts
online
Creation
Evaluation

Figure 3. The iterative cycle of DaPIS design

1. DaPIS version 1, in July-August 2017, at the Legal Design Lab of Stanford Law School (US): one exploratory workshop to design the first prototypes of the icon set;
2. DaPIS version 2, in March 2018, at the CIRSFID, University of Bologna, Italy, in collaboration with the Academy of Arts of Bologna and Società Italiana Informatica Giuridica: three multi-stakeholders' workshops focused on further icon design and redesign of DaPIS version 1;
3. DaPIS version 3, in July 2018, at the CIRSFID in collaboration with the Academy of Arts of Bologna: one last workshop to refine DaPIS version 2.

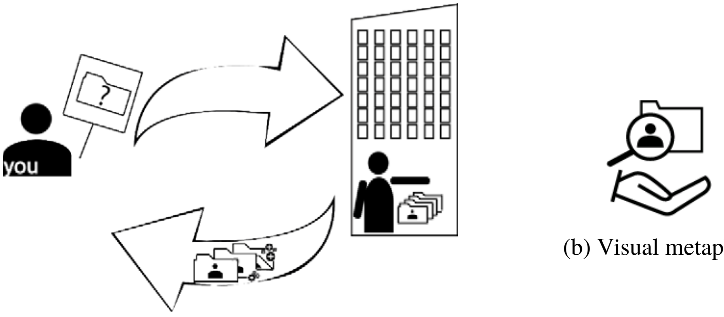
The first workshop was exploratory: it mainly served to indicate best practices for icon design of data protection concepts (i.e. mostly abstract and unfamiliar legal-technical concepts) and to rule out the less viable hypotheses that had been initially postulated. For instance, the modularity of composition of icons (see Section 4.4) was one of the proposed strategies that later informed the further development of the icon set. On the contrary, after the first icons' evaluation study and during the following workshops, the methodological choice to design literal and detailed representation of complex notions (e.g. the right to access) for alleged ease of interpretation was abandoned in favor of visual metaphors that could summarize the same concept in a reduced number of pixels and could be more easily used in responsive design (Figure 4).

During the following workshops, the missing visualizations for the ontological classes described above were systematically prototyped and vetted in a coherent manner with respect to the other existing icons; those icons that had shown major flaws during the user studies were re-elaborated; and, finally, for some concepts alternative solutions were conceived[8]. DaPIS has thus been designed in an iterative manner, through a continuous discussion, vetting and refinement of alternative ideas and prototypes during multiple cycles of evaluation and thanks to participants with composite backgrounds.

4.3. Addressing the Challenge from a Multi-Disciplinary Perspective

Diverse mental models and visual vocabularies derived from different backgrounds and experiences have been explicitly taken into consideration for the design of DaPIS. A pre-

[8] In [20], Chapter 5, thorough and extensive details about DaPIS design and evaluation are provided. See also: http://gdprbydesign.cirsfid.unibo.it/.



(b) Visual metaphor

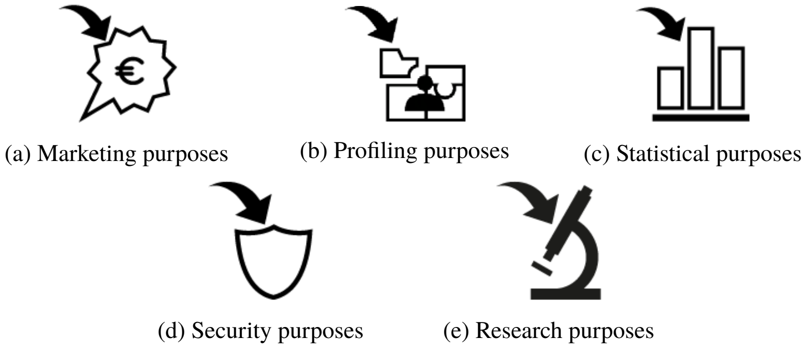
(a) Literal representation

Figure 4. Two subsequent versions of the right to access. On the left-hand side, the first prototype that attempts to literally reproduce the definition of the concept. On the right-hand side, the re-elaboration of the prototype into a less literal, but more contained representation

cise methodological choice was represented by the involvement of different stakeholders in the phase of creation of DaPIS in order to leverage their multiple skills and assets. Participatory design enables mutual learning among individuals with different mindsets and levels of expertise, make implicit, expertise-specific assumptions evident and debunk common domain-dependent misconceptions. Concretely, in this case, it meant that experts with a technical background clarified specialized notions such as encryption or pseudonymization to the other members of the group to facilitate the process of visualization; lawyers and legal scholars interpreted GDPR's definitions, provided appropriate examples and repeatedly voiced their concerns about misinterpretation and oversimplification, ultimately influencing icons' design; graphic designers and other professionals from visual disciplines provided the guidelines, the techniques and the tools to create appropriate visualizations for the intended audience and the intended medium; representatives of the businesses-world expressed the challenges for widespread adoption and effective implementation of the icon set; and finally laypeople added non-expert, but at the same time non-trivial, views and knowledge to the design process, for instance about the visual conventions they were familiar with [37]; [38]. Treating such a composite group of individuals as co-designers allowed for the creation of an icon set that epitomizes a synthesis of the different views, needs and concerns of the people that might be, in various manners, impacted by it.

4.4. Modularity and Compositionality of the Icons

The use of the ontology as conceptual framework oriented the icon design towards a compositional visual vocabulary: each graphical element corresponds to one and only one ontological element. Such basic elements can be combined to create more complex notions. For instance, the purposes of processing are represented by an arrow, while the concept of right is visualized with an upward-facing hand. Thus, the subclasses of purposes (Figure 5) and the subclasses of rights (Figure 6) contain such elements to indicate that they belong to the same class, but are complemented with other elements that specify their meaning and distinguish them from the icons of the same class. Similarly, personal data is represented as a prototypical file folder with a users' figure atop of it, while the outcome of processing operations on such data are visualized as variations of the folder (Figure 7). The complete icon set, composed of 33 icons, is freely downloadable[9] and is more thoroughly illustrated in [20].


(a) Marketing purposes

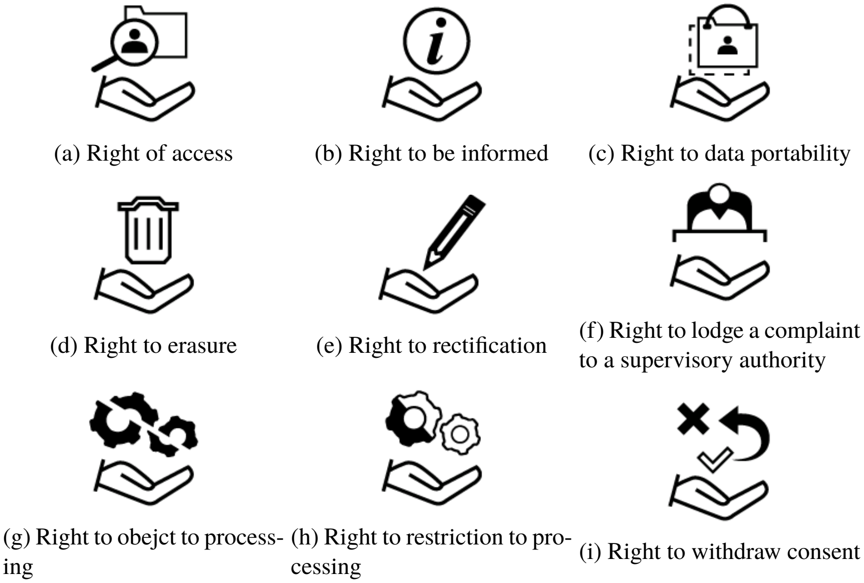
(b) Profiling purposes

(c) Statistical purposes

(d) Security purposes

(e) Research purposes

Figure 5. Icons representing the processing purposes: the recurrent arrow stands for 'purpose', while the complementary element specifies the type of purpose


(a) Right of access

(b) Right to be informed

(c) Right to data portability

(d) Right to erasure

(e) Right to rectification

(f) Right to lodge a complaint to a supervisory authority

(g) Right to obejct to processing

(h) Right to restriction to processing

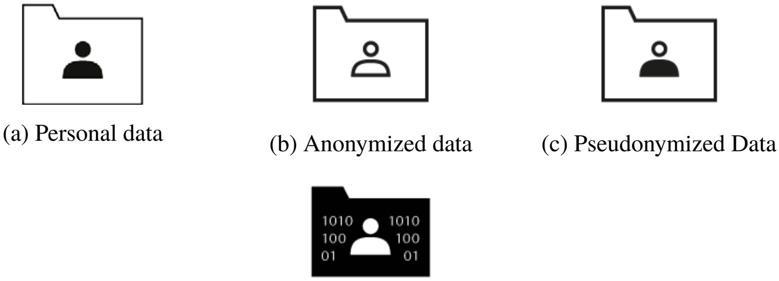
(i) Right to withdraw consent

Figure 6. Icons representing the rights of the data subject. The recurring element of the upward-facing hand symbolizes the 'right' and is meant to convey the concept of 'being in control' and 'having the power over' the element located above it, which specifies the meaning of the icon

4.5. Evaluation of the Icon Set

As described above, an evaluation phase followed every phase of major (re)design of DaPIS, for a total of three:

[9] http://gdprbydesign.cirsfid.unibo.it/dapis-2/.



(a) Personal data

(b) Anonymized data

(c) Pseudonymized Data

(d) Encrypted Data

Figure 7. The variations of the icons representing personal data as outcomes of different processing operations

1. Evaluation of DaPIS 1: carried out at Stanford (US); 16 participants; origin: mostly American; age: between 19 and 76 years old; level of education: mostly with a high school diploma;
2. Evlauation of DaPIS 2: carried out in Bologna (IT); 16 participants; origin: Italian; age: between 20 and 29 years old; level of education: mostly with at least a Bachelor's degree;
3. Evaluation of DaPIS 3: carried out in an online environment, on the research website; 10 participants; origin: Italy, Armenia, Iran, Canada and Greece; level of education: mostly with a Master's degree.

Main aim of such assessments was the evaluation of the iconographical choices made during the design workshops. Key dimensions that were considered were legibility (i.e. the ease of recognizition of the different elements composing each icon) and ease of understanding (i.e. the correct matching between graphical symbol and its meaning) [39]. For the latter, however, standard evaluation frameworks are either meant for those symbols whose referent is known to users (e.g. the concept of airplane)[10]; or for those symbols whose referent is unknown (e.g. the concept of pseudonymization) but a previous phase of familiarity training has been carried out[11]. Since none of such evaluation methods were appropriate for a one-time only assessment, a subjective estimation of goodness of fit between definition of a concept and the proposed graphical representation [40] and explanations thereof were asked, to provide an indication for those solutions considered more or less promising. In order to do so, even the degree of agreement among the respondents' answers was taken into consideration because indicative in this respect: great variation in the answers implies disagreement about the icons' efficacy, while greater uniformity indicates consensus and, hence, pinpoints the more easily recognizable icons. For instance, positive ratings coincided for the icon representing the transfer to countries outside of the EU because it uses an easily recognizable prototypical representation (i.e. the stars in circle of the EU flag and the personal data folder), while diverse degrees of appreciation were gathered by, e.g., the icon representing public interest. This variety of ratings is also due to the fact that there are some graphical elements are more familiar to users than other, while concrete objects are also more easily visualizable,

[10] ISO (2014). *ISO 9186-1:2014. Graphical Symbols – Test Methods – Part 1: Method for Testing Comprehensibility.*

[11] ISO (2014). *ISO 9186-3:2014. Graphical Symbols – Test Methods – Part 3: Method for Testing Symbol Referent Association.*

and thus, apprehended. The motivations and explanations for the ratings provided by the participants constituted a valuable informative feedback that was crucial for further icon re-elaboration.

5. Conclusions and Future Work

This Chapter has briefly presented the development and the main outcomes of a multidisciplinary project that combines machine-readable representations of legal information with corresponding (still underresearched, but increasingly investigated) human-oriented visual representations. The case study analyzed in these pages concerns the data protection domain and is meant to propose an implementation of the transparency principle set forth by the General Data Protection Regulation. In particular, an overview of PrOnto, an ontology of the GDPR's concepts (and norms), and DaPIS, a data protection icon set modeled on such concepts where provided. The explanation of the design and evaluation of such icon set has constituted the focus of this Chapter.

Future research should proceed in multiple directions. The PrOnto ontology is currently being developed, refined and expanded. Appropriate icons for additional classes can be created or mutuated from other icon sets that have been developed or are currently under development. However, it should be researched the number of icons that an individual can learn and retain without feeling overwhelmed. Furthermore, notwithstanding the user studies briefly described above, there are limitations that should be addressed in the future. Firstly, the participants to the studies were few, very diverse in terms of origin but uniform in terms of educational level (i.e. medium-high). Some of them had legal expertise or strong familiarity with technologies, while others did not. Secondly, since one of the main obstacles to ease of recognition is the lack of familiarity with the graphical symbol or with its referent, it should be expected that the effect of training increases recognition rates and determines easier recall. Thus, longitudinal studies that envisage consequent phases should be preferred [41] and should carried out involving individuals coming from all the Member States and of more varied demographics. Finally, as illustrated in [21], DaPIS should also be judged with respect to its function in context: namely, the capacity of its icons to act as information markers that effectively support the navigation through large amounts of legal information and increase speed and accuracy of comprehension.

DaPIS does not aim to be the ultimate, immutable data protection icon set to be standardized across the European member states. There even are some icons that would benefit from a re-design and better performing alternative solutions could be advanced. Other projects having similar aims exist [34], but might reach different outcomes. This research rather details a methodology for a design process that, on the one hand, values the multidisciplinarity that is necessary when dealing with complex contemporary topics like device-mediated data processing. On the other hand, it strives to an effective integration with semantic technologies for legal data that are mostly concerned with machine-interpretable meanings, but more rarely remember the human end-user. The goal of this research is to possibly inform the preparatory work and the decisions of the European Commission about the creation and implementation of the GDPR's icons and to stimulate an international, interdisciplinary debate.

References

[1] Nissenbaum, H. (2011). A Contextual Approach to Privacy Online. *Daedalus, 140*(4), 32-48.

[2] Monteleone, S. (2015). Addressing the Failure of Informed Consent in Online Data Protection: Learning the Lessons from Behaviour-aware Regulation. *Syracuse Journal of International Law and Commerce, 43*, 69.

[3] Schaub, F., Balebako, R., Durity, A. L. & Cranor, L. F. (2015). A Design Space for Effective Privacy Notices. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*, 1-17.

[4] Solove, D. J. (2013). Privacy Self-Management and the Consent Dilemma. *Harvard Law Review*, *126*, 1880.

[5] Oehler, A. & Wendt, S. (2017). Good Consumer Information: The Information Paradigm at Its (Dead) End? *Journal of Consumer Policy, 40*(2), 179-191.

[6] Berger-Walliser, G., Bird, R. C. & Haapio, H. (2011). Promoting Business Success through Contract Visualization. *Journal of Law, Business & Ethics, 17*, 55.

[7] Robinson, N., Graux, H., Botterman, M. & L. Valeri (2009). *Review of the European Data Protection Directive (sponsored by the Information Commissioner's Officer)*. RAND Europe.

[8] Ben-Shahar, O. & Schneider, C. E. (2014). *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton University Press.

[9] Fabian, B., Ermakova, T. & Lentz, T. (2017). Large-scale Readability Analysis of Privacy Policies. In *Proceedings of the International Conference on Web Intelligence*. ACM, 18-25.

[10] Article 29 Data Protection Working Party (2018). *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679*, 17/EN WP251rev.01.

[11] Haapio, H., Plewe, D. & deRooy, R. (2017). Contract Continuum: From Text to Images, Comics, and Code. In Schweighofer, E., Kummer, F., Hötzendorfer, W. & Sorge, C. (Eds.). *Trends and Communities of Legal Informatics. Proceedings of the 20th International Legal Informatics Symposium IRIS*. OCG Verlag, 411-418.

[12] Palmirani, M. & Vitali, F. (2011). Akoma-Ntoso for Legal Documents. In Sartor, G., Palmirani, M., Francesconi, E. & Biasiotti, M. A. (Eds.). *Legislative XML for the Semantic Web*. Springer, 75-100.

[13] Passera, S. (2015). Beyond the Wall of Text: How Information Design Can Make Contracts User-friendly. In *International Conference of Design, User Experience, and Usability*. Springer, 341-352.

[14] Waller, R., Waller, J., Haapio, H., Crag, G. & Morrisseau, S. (2016). Cooperation through Clarity: Designing Simplified Contracts. *Journal of Strategic Contracting and Negotiation, 2*(1-2), 48-68.

[15] Haapio, H. & Hagan, M. (2016). Design Patterns for Contracts. In Schweighofer, E., Kummer, F., Hötzendorfer, W. & Borges, G. (Eds.). *Networks. Proceedings of the 19th International Legal Informatics Symposium IRIS*. OCG Verlag, 381-388.

[16] Haapio, H. & Passera, S. (2019). Contracts As Interfaces: Exploring Visual Representation Patterns in Contract Design. In Katz, D., Bommarito, M. & Dolin, R. (Eds.). *Legal Informatics*. Cambridge University Press.

[17] Haapio, H., Hagan, M., Palmirani, M. & Rossi, A. (2018). Legal Design Patterns for Privacy. In Schweighofer, E., Kummer, F., Saarenpää, A. & Schafer, B. (Eds.). *Data Protection/LegalTech. Proceedings of the 21th International Legal Informatics Symposium IRIS*. Editions Weblaw, 445-450.

[18] Rossi, A., Ducato, R., Haapio, H. & Passera, S. (2019). When Design Met Law: Design Patterns for Information Transparency. *Droit de la Consommation = Consumenterecht: DCCR*, 122-123, 79-121.

[19] Rossi, A., Ducato, R., Haapio, H., Passera, S. & Palmirani, M. (2019). Legal Design Patterns: Towards A New Language for Legal Information Design. In Schweighofer, E., Kummer, F. & Saarenpää, A. (Eds.). *Internet of Things. Proceedings of the 22nd International Legal Infomatics Symposium IRIS*. Editions Weblaw, 517-526.

[20] Rossi, A. (2019). *Legal Design for the General Data Protection Regulation. A Methodology for the Visualization and Communication of Legal Concepts*. PhD Thesis, Alma Mater Studiorum Università di Bologna, PhD in Law, Science and Technology (forthcoming).

[21] Palmirani, M., Rossi, A., Martoni, M. & Hagan, M. (2018). A Methodological Framework to Design a Machine-readable Privacy Icon Set. In Schweighofer, E., Kummer, F., Saarenpää, A. & Schafer, B. (Eds.). *Data Protection/LegalTech. Proceedings of the 21th International Legal Informatics Symposium IRIS*. Editions Weblaw, 451-454.

[22] Monica Palmirani, M., Sperberg, R., Vergottini, G. & Vitali, F. (2016). *Akoma Ntoso Version 1.0 Part 1: XML Vocabulary*. OASIS Committee Specification Draft 02 / Public Review Draft 02. Technical report.

[23] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C. & Robaldo, L. (2018). PrOnto: Privacy Ontology for Legal Reasoning. In *eGovis 2018, the 7th International Conference on Electronic Government and the Information Systems Perspective*. Springer, 139-152.

[24] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C. & Robaldo, L. (2018). PrOnto: Privacy Ontology for Legal Reasoning. In *Proceedings of the 18th European Conference on Digital Government*. Academic Conferences and Publishing International Limited, 142-151.

[25] Palmirani, M., Martoni, M., Rossi, A., Bartolini, C. & Robaldo, L. (2018). Legal Ontology for Modelling GDPR Concepts and Norms. In Palmirani, M. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 31st Jurix Conference*. IOS Press, 91-100.

[26] Article 29 Data Protection Working Party (2017). *Guidelines on Personal Data Breach Notification Under Regulation 2016/679*, 18/EN WP250rev.01.

[27] Barabucci, G., Di Iorio, A., Poggi, F. & Vitali, F. (2013). Integration of Legal Datasets: From Metamodel to Implementation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. ACM, 585.

[28] Functional Requirements for Bibliographic Records (2009). *International Federation of Library Associations and Institutions*. Technical report.

[29] Breuker, J. A. P. J., Hoekstra, R., van den Berg, K., Rubino, R., Sartor, G., Palmirani, M., Wyner, A. & Bench-Capon, T. (2007). *OWL Ontology of Basic Legal Concepts (LKIF-Core). Estrella: Deliverable*. UVA. Technical report.

[30] Gangemi, A., Peroni, S., Shotton, D. & Vitali, F. (2017). The Publishing Workflow Ontology (PWO). *Semantic Web, 8*(5), 703-718.

[31] Peroni, S., Palmirani, M. & Vitali, F. (2017). UNDO: The United Nations System Document Ontology. In *International Semantic Web Conference*. Springer, 175-183.

[32] Guarino, N. & Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM, 45*(2), 61-65.

[33] Graf, C., Hochleitner, C., Wolkerstorfer, P., Angulo, J., Fischer-Hübner, S. & Wästlund, E. (2011). *Final HCI Research Project*. PrimeLife Consortium. Technical report.

[34] Privacy Tech. *Privacy Icons*. https://www.privacytech.fr/privacy-icons/.

[35] Hagan, M. (2017). *Law by Design*, http://www.lawbydesign.co.

[36] Van Der Velden, M. & Moertberg, C. (2015). Participatory Design and Design for Values. In *Handbook of Ethics, Values and Technological Design*. Springer, 41-66.

[37] Rossi, A. & Palmirani, M. (2018). From Words to Images Through Legal Visualization. In *AI Approaches to the Complexity of Legal Systems*. Springer.

[38] Rossi, A. & Haapio, H. (2019). Proactive Legal Design: Embedding Values in the Design of Legal Artefacts. In Schweighofer, E., Kummer, F. & Saarenpää, A. (Eds.). *Internet of Things. Proceedings of the 22nd International Legal Infomatics Symposium IRIS*. Editions Weblaw, 537-544.

[39] Dewar, R. (1999). *Visual Information for Everyday Use: Design and Research Perspectives*, Chapter Design and Evaluation of Public Information Symbols. Taylor and Francis, 285-303.

[40] Wogalter, M. S., Silver, N. C., Leonard, S. D. & Zaikina, H. (2006). *Handbook of Warnings*, Chapter Warning symbols. Lawrence Erlbaum Associates Mahwah, 159-176.

[41] Flick, U. (2018). *An Introduction to Qualitative Research*. Sage Publications Limited.

This page intentionally left blank

Part III

Experiences, Good Practices and Critical Issues

This page intentionally left blank

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190022

Legal Information Institutes and AI: Free Access Legal Expertise

Graham GREENLEAF, Andrew MOWBRAY and Philip CHUNG
Australasian Legal Information Institute - AustLII (Australia)

Abstract. The use of Artificial Intelligence (AI) in law has again become of great interest to lawyers and government. Legal Information Institutes (LIIs) have played a significant role in the provision of legal information via the Web. The concept of 'free access to law' is not static, and its principles now require a LII response to the renewed prominence of AI, possibly to include improving and expanding free access to legal advice. This overview of one approach, from justification to implementation, considers the potential for AI-aided free legal advice, its likely providers, and its importance to legal professionalism. The constraints that 'free' imposes lead to the potential roles LIIs may realistically play, and suggested guidelines for development of sustainable systems by free access providers. The AI-related services and tools that the Australasian Legal Information Institute (AustLII) is providing (the 'DataLex' platform) are outlined. Finally, ethical (or governance) issues LIIs need to address are discussed.

Keywords. artificial intelligence, LII - Legal Information Institute, legal advisory system

The use of Artificial Intelligence (AI) in law, including in relation to decision-support systems, has again become a matter of great interest to both the legal profession and to government. The previous wave of enthusiasm for, and investment in, 'AI and law' from the early 1980s to the mid-1990s was to a large extent supplanted by the development of the World-Wide-Web and the provision of legal information via the Web. Legal Information Institutes (LIIs) and the Free Access to Law Movement (FALM) played a very significant role in those developments [1]. What roles might LIIs play in this new AI-oriented environment?

The concept of 'free access to law' is not static, and has evolved over the past quarter-century [2]. The principles of free access to law now require a LII response to the renewed prominence of AI-related developments in law, which could include improving and expanding free access to legal advice, as part of 'free access to law', consistent with those of FALM's *Declaration of Free Access to Law*[1]. 'Freeing the law' is a continuous process.

This Chapter provides an overview of one approach to all aspects of this question, from justification to implementation. We commence with a discussion of the potential for provision of AI-aided free legal advice, its likely providers, and its importance to the future of the legal profession. We then consider the constraints that the requirement of 'free' imposes, including on what types of free legal advice systems are sustainable, and what roles LIIs may realistically play in the development of such a 'commons of free

[1] http://www.falm.info/declaration/.

legal advice'. We suggest guidelines for development of such systems. The AI-related services and tools that the Australasian Legal Information Institute (AustLII) is providing (the 'DataLex' platform) are outlined, including how they implement these guidelines. Finally, we suggest questions concerning ethical (or governance) principles LIIs need to address when they are involved in using AI tools.

1. The New Threats and Promises that AI Presents to Legal Advice

The 'Web 2.0' context since about 2004[2] creates a very different environment from the pre-1995 (pre-Internet, in popular usage) context of the first wave of 'AI and law' This context makes it more feasible to talk about the collaborative development of free legal advice services based on AI. The reasons include the significant roles that FOSS (free and open source software) and open content (exemplified by Creative Commons licensing and Wikipedia) have had on the development of the Internet; the much greater sophistication of interfaces; and the differences that interaction between AI-based tools and huge amounts of free access legal content can make.

1.1. The Trajectory of Digitisation of Legal Information Toward a Commons

We can distinguish three types of digitisation relevant to the giving of professional legal advice: representation of information used by experts; representation of expertise and its general application; and application of expertise to individual situations. These categories overlap in reality, these distinctions enable us to consider more precisely [3] how likely is it that each category will be 'liberated' and become part of the commons (in which we include availability for free access).

(I) Representing Expert Domain Information – 'Raw' (primary) information used by experts is the most likely aspect of expertise both to be digitised and to become part of the commons. Databases of primary information essential to legal professionals (legislation, treaties, court decisions etc.) are already substantially digitised and available online, and with increasing utility (e.g. smarter retrieval systems, and smarter data structures). In many countries substantial amounts are available as commons, at least for free access and often as open content, usually via government sources. In a few dozen countries such as Australia, free access 'legal information institutes' (LIIs) aggregate this data and add value to it, making it a resource used by professionals and the general public alike. Even there, some primary information is only available commercially, include standards, 'authorised' reports (monopolistic practices arising from privileged citation practices), and important 'pre-LII') historical data. However, in less than 25 years since the start of widespread availability of such data via the Web, the increase in free availability is extraordinary, and is tending toward a comprehensive commons.

(II) Representing expertise in general form – When professional expertise is represented (or embodied or reified) this is usually in a generalised form which may or may not be applicable to an individual situation where expertise is needed, because of the enormous variation of individual situations which may arise. It is up to the reader (usu-

[2] The term 'Web 2.0' was popularised from around 2004, generally taken to include web services catering for user-generated content (including all social media, blogs, Twitter etc.), and many software enhancements allowing much more responsive interfaces than plain HTML.

ally the correct term) to apply the expertise to the individual situation. Legal professionals represented their expertise in many ways prior to the Internet – in textbooks, journal articles, encyclopedias, and in very significant, but more mundane, forms such as citators and checklists (often as supervisors of non-professionals).

In the pre-Internet era, compilations of expertise may have been collective (e.g. commissioned encyclopedia articles, or *Halsbury's Laws*), but were very rarely 'crowd sourced'. The economics of publishing meant that such reification of expertise could rarely be provided as a commons, and instead it usually became an economic asset of a commercial publisher and an author. The Internet changes some but not all of these factors. Expertise is a very valuable asset of many professionals. It is very time-consuming to consciously embody it in any form, and many professionals are very reluctant to 'give it away', either because they believe it gives them a competitive advantage, or because they would prefer to be paid by a publisher, or because publishing expertise is time-consuming, difficult and potentially risky. Commercial publishers of such expertise will not disappear. Commons must always coexist with commerce.

However, the last quarter century has brought many changes, the revolutionary potential of which are only becoming apparent through the accretion of successes, including free access repositories of current scholarship, archives of published journals, and changing academic funding requirements. The crowd-sourced Wikipedia demonstrates that under certain circumstances (including viral licensing), the expert and non-expert public can combine to create the largest, free, and probably by now most reliable encyclopedia. However a 'closed wiki' model, where content may only be edited by professionals may be more suitable for law, because of its emphasis on authority. Successful commons examples exist including multi-author guidebooks [4], and automated citators performing to professional levels [5]. The result is that the combination of factors such as these – peer-reviewed free content; funding body pressure; viral licensing; crowd-sourcing; collaborative editing by closed professional groups; and automated substitutions for expertise – and many others, may threaten the viability of some types of commercial control of the publishing of expertise. More importantly, they demonstrate that it is becoming viable for professionals to control the representation of their own expertise, as a commons.

(III) Applying Expertise to Individual Situations – It is the third category, the application of expertise to individual situations (the problems of individual clients) via programs, which is seen widely as a major threat to the future of professionals and professions [4]. At present, the number of convincing examples and their commercial viability do not make it inevitable that there will be generalised dire results for professions. To understand the likely implications, it is necessary to distinguish at least three types of the programmatic applications of legal expertise: human expertise embodied in knowledge-bases which interact with programs; embedded knowledge in artifacts; and machine-generated expertise. The first is most relevant to free legal advice providers. The question is whether, in those areas where legal expertise can be effectively captured in knowledge-bases to be used in decision-support systems, can they be developed as a commons, or only as commercial products? The following sections explore this further.

1.2. AI and Threats to Legal Professionalism

It is somewhat ironic that one of the arguments favouring the development of a commons of legal expertise is the threat that the application of AI to law poses to the legal profession and individual legal professionalism. The extent of the threat (also presented as a promise of efficiency) is still very difficult to estimate, though often claimed to be extreme. The threat has three main sources. Where there is a substantial market for solutions to a category of legal problem, expert advisory systems can economically automate answering problems up to a certain level of complexity. Beyond that, knowledge acquisition and other bottlenecks make their applicability unproven. Embedded knowledge delivered through software will continue to permeate the material world and to impose 'normal' behaviour which adheres to medical, accounting or legal norms (including some 'smart contracts'). Machine-generated expertise will be relied upon increasingly in relation to the set of problems where prediction of a 'correct' answer is sufficient (by whatever path it is reached), and explanations in terms of underlying causes and human reasoning are not required ([4], 2.3).

If we assume that these nascent developments will accelerate, what are the implications for legal professionals? Instead of a relatively prestigious and well-paid class of 'para-professionals', who support the delivery of semi-automated packaged commercial services, there may instead develop an intermediate category of what we could call 'pseudo-lawyers', who have the training, the formal status, and the self-image of a lawyer, but are really closer to a low-paid paralegal doing repetitive work involving moderate levels of expertise. This will usually involve driving and interpreting computerised products developed by those with more expertise. Another set of possibilities (various 'Uber models') involves teams of individual legal providers put together on an ad hoc basis by intermediaries ('platform providers'), likely to involve the platform provider taking a major share of the funds paid for provision of services.

In any of these future scenarios for individual lawyers, who will own the intellectual property in the software and applications used in these types of provisions of services? It is unlikely to be the employed 'pseudo-lawyers' or the service provider in an 'Uber model': the large firm employer, or the provider of the platform will be likely to develop such expensive tools themselves, or will have the necessary very significant funds to buy them (and keep them updated) from a large commercial legal publisher. The employed solicitor, the small practitioner or the barrister in chambers, except those at the higher levels of the profession, will not be able to afford the modern equivalents of legal professional tools. We argue that protecting their professionalism may also help to produce free legal advice services.

2. An Alternative Future: A Commons of Legal Expertise

Although there is as yet no obvious tendency toward commons in relation to the three categories of software-based application of expertise to individual cases, we argue that this can be encouraged to develop. Tools for knowledge engineering and for creating machine-generated expertise are available as FOSS and are of high quality, but the communities of users necessary to develop applications (similar to the FOSS or Wikipedia communities) have not yet developed. The employed solicitor, small practitioner or barrister is unlikely to contribute spontaneously to the development of commons. But the risk for such professionals in not having any role in the development of AI tools in law is that they will lose control of their standing, abilities and future as professionals, to a far greater extent than in the pre-AI structures of the legal profession.

The alternative is that there be at least some part of the development and use of AI in law that is open to participation by any lawyers, and which collectively may provide a set of AI-based applications that are an alternative to those controlled by mega-firms in law and consultancy, and the oligopoly of large publishers. For sole practitioners and small firms, some such collaboration many be the only strategy possible for them to participate.

From where could such a collaborative alternative arise? We argue that it could arise primarily from those organisations that seek to provide free legal advice, and be driven largely by their needs, but could expand to involve other participants in the legal profession.

2.1. The Providers and Constraints of Free Legal Advice

There are many situations where, at least in a country like Australia, our social expectation is that legal advice be provided without cost to the public, whether as consumers, citizens or (sometimes) litigants. The organisations most likely to be involved in providing such free legal advice are quite diverse, and include government legal aid providers, community legal centres, government and community consumer advice centres, specialist NGOs in law-related areas, government agencies giving advice relevant to their functions, and ‘chamber magistrates’ in courthouses. The legal profession, through state and regional Law Societies and advice centres they provide, and through the extensive *pro bono* schemes, also contributes. University law schools, through their involvement in community legal centres and internships in other organisations, are potential sources of contributors who often have high computing skills. Bodies assisting the legal profession as a whole to avoid liability problems, such as some legal insurers, might also wish to participate.

A common factor in most of these providers of free legal advice is that, if they choose to develop AI-related tools to assist their work, they will usually have to do so within very constrained development and maintenance budgets for software or applications. They are not in a position to pass on such costs to clients, or to purchasers of applications. Government or other grants for such developments may provide up-front development costs (at least while the hype cycle for AI is rising) but will rarely cover ongoing maintenance for applications as the law changes, or technical issues arise. Bringing in out-of-house consultants on specialised software problems, or as ‘knowledge engineers’ in relation to particular legal domains, is likely to be very expensive. It is therefore a reasonable assumption that, at least in the medium to long term, providers of free legal advice will have to work within significant financial constraints that are more severe than those experienced by commercial providers.

The implications of these constraints – limited institutional range of providers, and limited financial resources – affect the types of legal advisory systems that it is practical for this sector to develop and support.

2.2. Free Legal Advisory Systems: Guidelines for Sustainability

We have previously set out and justified our views on what approach to the use of AI tools is most likely to be of value to a free legal advice service ([6], 3.1-3.16). These guidelines are based on the assumptions discussed above of the likely limited financial and personnel resources of such a service, and on our own lengthy experience with the DataLex project. They are implemented in the DataLex platform discussed in Section 3.

First, the ‘AI and law’ systems that such a service could be expected to find useful are those that justify their answers at least in part in terms of the formal sources of law. These constraints will mean that only some types of ‘AI and law’ tools are suitable to their needs.

Second, looked at from the user perspective, which could be that of an employee of a free legal advice service, or perhaps one of its clients, what counts as a useful level of legal expertise is relative. A system may be valuable to a class of users even though it has a relatively low point at which it admits that a problem is beyond its expertise, and it may serve as a method of triage. In any event, it is not realistic to try to build legal expert systems that encapsulate all the knowledge necessary to answer user problems. The more realistic aim is to build decision support systems, in the use of which the program and the user in effect pool their knowledge/expertise to resolve a problem. Expertise can and should be represented and utilised by programs in many ways. This means the knowledge-based system (the knowledge representation and the program) should not be ‘closed’: it must be integrated with text retrieval, hypertext and other tools which allow and assist the user to obtain access to whatever source materials are necessary to answer the parts of a problem dependent on the user’s expertise. The result is an integrated decision-support system.

Third, looked at from the developer perspective, the key contextual factor is that user-organisations such as free legal advice services, will probably need to both develop and maintain their own knowledge-bases, as the only available domain experts. The systems which non-technical legal domain experts are most likely to be able to develop and maintain are those which represent legal knowledge in a way which has a reasonably high level of isomorphism (one-to-one correspondence) with the legal sources on which it is based, where the representation is reasonably close to natural language, and where it is not necessary to prescribe the order(s) of the procedural steps necessary to reach a solution to a problem, but only to declare what legal knowledge is available, and leave it to the system to undertake the steps to apply that knowledge.

Fourth, correctly choosing the type of problem where ‘AI and law’ techniques are most likely to be appropriate is essential. Problem areas based on legislation, or procedural steps, and where there is complexity, will probably give the best results. Problems involving multiple instances of one factor increase logical difficulty. If it is administratively possible to have multiple organisations collaborate to build and maintain a legal knowledge-base, this may increase sustainability.

Is the approach sketched under these four headings out-of-step with current approaches to the use of AI in legal applications? Ashley, a leading current proponent of the field of AI and law, might well regard our ambitions for system development as unnecessarily modest (or perhaps just the product of our constraining assumptions), but there is little that is inconsistent between this ‘legal decision support system’ approach taken by the DataLex Project and its underlying rationales, and the ‘cognitive computing’ paradigm he advances [7].

2.3. The Likely Roles of LIIs

Fifth, we conclude that free access legal information institutes (LIIs) are unlikely to be the builders of legal knowledge-bases in particular legal domains, because they do not have the necessary in-house expertise in legal subject domains. They have neither the client-base that provides a continuing need for such expertise, nor the funds to retain such expertise from outside (at least not on a continuing basis, beyond an initial grant). As a result, LIIs are much more likely to be the providers of tools by which such knowledge-bases are built, the free access legal infrastructure within which they are built, and education and support for those organisations that use their tools and services to build and maintain subject-area applications. In light of that conclusion, we now move to the tools and services that AustLII is building.

3. AI in a LII: AustLII's DataLex Implementation

The Australasian Legal Information Institute (AustLII), through its DataLex project [6]; [8] is developing tools and infrastructure so as to implement the above 'sustainable legal advisory systems' approach to AI and law in the context of a LII. This platform includes five main elements, rectangles in the following diagram. The features of each are then summarised (Figure 1).

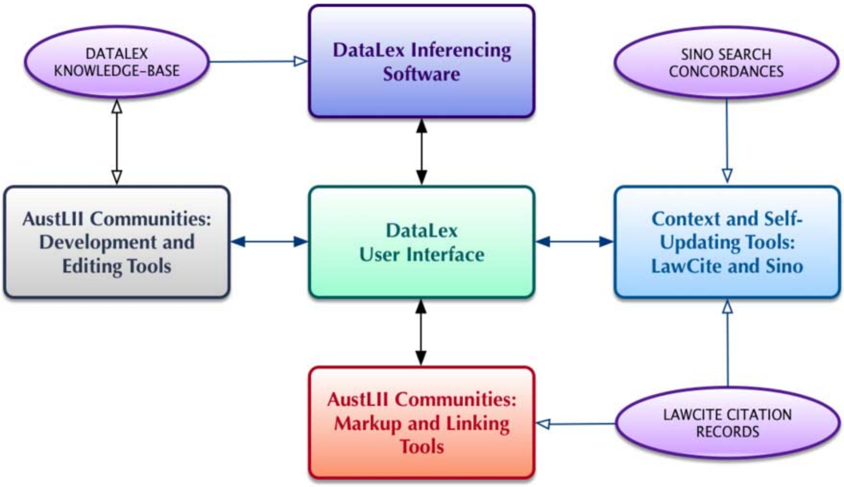

DATALEX KNOWLEDGE-BASE
DataLex Inferencing Software
SINO SEARCH CONCORDANCES
AustLII Communities: Development and Editing Tools
DataLex User Interface
Context and Self-Updating Tools: LawCite and Sino
AustLII Communities: Markup and Linking Tools
LAWCITE CITATION RECORDS

Figure 1. Elements of AustLII's DataLex legal inferencing platform

3.1. The DataLex Inferencing Software

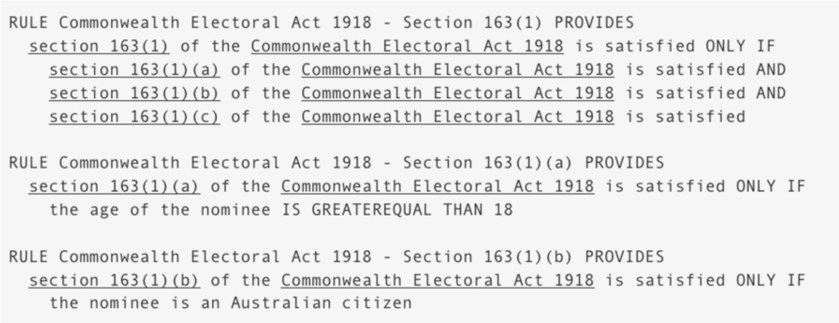
The DataLex inferencing software[3] primarily carries out rule-based reasoning. It has the following key features:

- Support for backward-chaining and forward-chaining rule-based reasoning. Rules are expressed in a declarative form.
- Rule-based reasoning is supplemented by procedural code, where procedural steps in reasoning are needed.

[3] The DataLex inferencing software was originally written by Andrew Mowbray, as y-sh ('y-shell'), with subsequent further layers by various authors including Simon Cant and Philip Chung, to enable web-based operation.

- Rule based reasoning is also supplemented by example-based (or ‘case-based’) reasoning[4], where needed.
- Rules of any degree of complexity may be written, using propositional logic.
- A quasi-natural-language knowledge-base syntax (ie one resembling English as far as is possible) is used to declare rules (and examples).
- There is no separate coding of questions, explanations and reports, because they are all generated automatically from the declared rules, in dialogues generated ‘on the fly’ when the system is in operation. This default operation can be customised where special circumstances require.
- Isomorphic (one-to-one) relationships between the knowledge-base and legislation is facilitated, and assists in debugging and updating.
- The previous three elements allow easier development, de-bugging and maintenance by domain experts (lawyers), without involvement by software experts or ‘knowledge engineers’.
- Collaborative development of larger applications across distributed knowledge-bases is supported.

An extract from the ElectKB knowledge-base [10] is shown in Figure 2:



```
RULE Commonwealth Electoral Act 1918 - Section 163(1) PROVIDES
  section 163(1) of the Commonwealth Electoral Act 1918 is satisfied ONLY IF
    section 163(1)(a) of the Commonwealth Electoral Act 1918 is satisfied AND
    section 163(1)(b) of the Commonwealth Electoral Act 1918 is satisfied AND
    section 163(1)(c) of the Commonwealth Electoral Act 1918 is satisfied

RULE Commonwealth Electoral Act 1918 - Section 163(1)(a) PROVIDES
  section 163(1)(a) of the Commonwealth Electoral Act 1918 is satisfied ONLY IF
    the age of the nominee IS GREATEREQUAL THAN 18

RULE Commonwealth Electoral Act 1918 - Section 163(1)(b) PROVIDES
  section 163(1)(b) of the Commonwealth Electoral Act 1918 is satisfied ONLY IF
    the nominee is an Australian citizen
```

Figure 2. Extract from the ElectKB knowledge-base

3.2. The AustLII Communities Environment – Integrating AI with a LII

The AustLII Communities environment is used to link automatically both knowledge-bases under development, and advisory systems when in operation, with all of the free access legal materials provided by a LII. The hypertext links in the above know-ledge-base extract are inserted automatically, using AustLII’s *findacts* software, into the knowledge-base as it is written and saved. Further examples of links from applications in operation are given below.

3.3. The DataLex Knowledge-base Development Tools

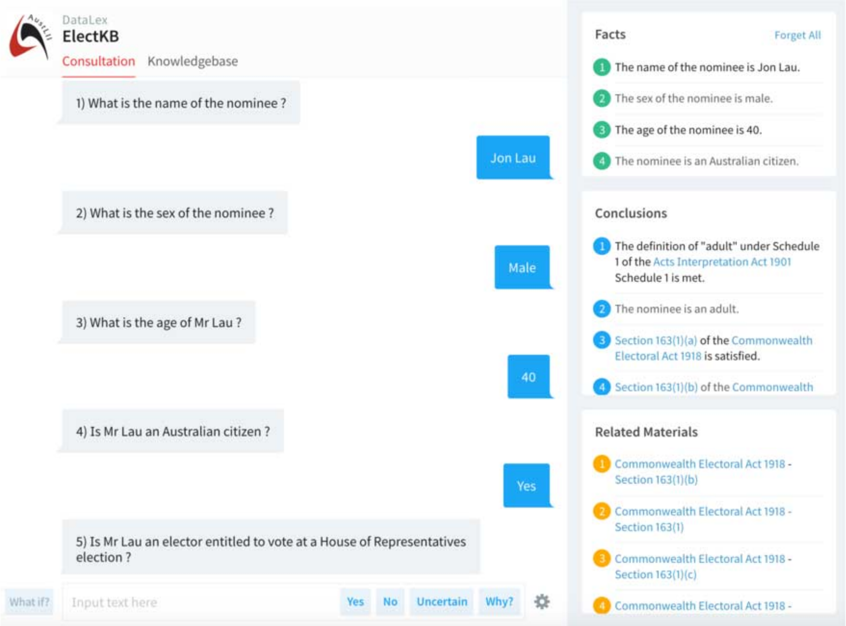
The DataLex development tools [11] are situated within the AustLII Communities infrastructure. They use a familiar wiki-like editing interface for development and maintenance of knowledge-bases (KBs). Development is within a closed wiki environment.

[4] PANNDA (Precedent Analysis by Nearest-Neighbour Discriminant Analysis); see [9] for details about the FINDER (finders’ cases) application of PANNDA.

3.4. The DataLex User Interface

The DataLex user interface uses the DataLex software and knowledge-bases, the linkages provided by the Communities environment, and user input, to provide legal advisory systems in operation. From Figure 3 the following screen, it can be seen that some of the features of the interface include:

- Questions, Facts, Conclusions, and Reports are all generated from the knowledge-base and user-provided facts, in understandable form, and are available on screen at all times.
- Facts can be deleted ('Forget?'), and questions than re-asked; Conclusions can be explained ('How?'); and reasons for Questions requested ('Why?'), generated in the same manner.
- The system also uses all information available to it, from the knowledge-base and user-supplied facts, to suggest other relevant Related Materials.



1) What is the name of the nominee ?

Jon Lau

2) What is the sex of the nominee ?

Male

3) What is the age of Mr Lau ?

40

4) Is Mr Lau an Australian citizen ?

Yes

5) Is Mr Lau an elector entitled to vote at a House of Representatives election ?

Facts

1. The name of the nominee is Jon Lau.
2. The sex of the nominee is male.
3. The age of the nominee is 40.
4. The nominee is an Australian citizen.

Conclusions

1. The definition of "adult" under Schedule 1 of the Acts Interpretation Act 1901 Schedule 1 is met.
2. The nominee is an adult.
3. Section 163(1)(a) of the Commonwealth Electoral Act 1918 is satisfied.
4. Section 163(1)(b) of the Commonwealth

Related Materials

1. Commonwealth Electoral Act 1918 - Section 163(1)(b)
2. Commonwealth Electoral Act 1918 - Section 163(1)
3. Commonwealth Electoral Act 1918 - Section 163(1)(c)
4. Commonwealth Electoral Act 1918 -

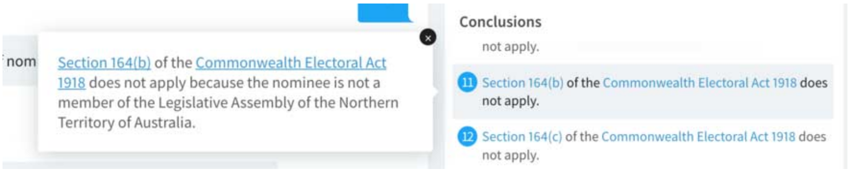
Figure 3. DataLex user interface features: 'Consultation', 'Facts'', 'Conclusions', 'Related Materials'

As the consultation continues, conclusions are shown on the right-hand side. Selection of a numbered conclusion results in a 'How' explanation of that conclusion being presented (Figure 4).

At the end of the consultation, a composite explanation of the final result, and of all the steps necessary for it to be reached, is displayed and may be exported to word processing or other programs for use.

3.5. The LawCite Citator and SINO Search Engine – Updating and Expanding Advice

SINO is the open source search engine, developed by AustLII [12], used to operate AustLII and other LIIs. The LawCite citator [6] is an automated international citator for



nom

Section 164(b) of the Commonwealth Electoral Act 1918 does not apply because the nominee is not a member of the Legislative Assembly of the Northern Territory of Australia.

Conclusions

not apply.

11 Section 164(b) of the Commonwealth Electoral Act 1918 does not apply.

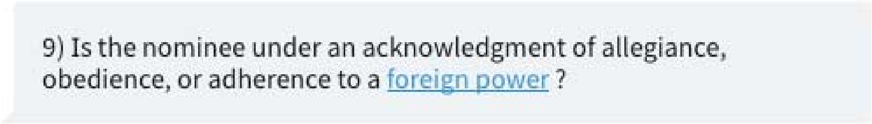
12 Section 164(c) of the Commonwealth Electoral Act 1918 does not apply.

Figure 4. DataLex ‘How’ explanation of a conclusion during consultation

case law and legal scholarship, accessible to end-users free of any user charges. It is developed and maintained by AustLII in conjunction with a consortium of participating legal information institutes (LIIs). LawCite currently contains index records of the citation histories of almost 5.7 million cases, law journal articles, law reform documents and treaties, going back to the 1300s. It includes citation records in significant numbers from court decisions in 75 countries. It is integrated fully into the operations of AustLII and other LIIs that use it. The technical details of LawCite are explained elsewhere [13].

The significance of both LawCite and SINO within the DataLex project is that they provide a means of (in effect) expanding the scope of a knowledge-base by providing users with access to knowledge which is not yet encoded within the knowledge-base. Examples are as follows, from the ElectKB knowledge-base [10] concerning disqualification for eligibility for election to the Australian federal Parliament:

1. Wherever the term ‘foreign power’ appears in a consultation dialogue, it does so as a hypertext link (Figure 5) which triggers a search over AustLII for all occurrences of ‘foreign power’ in the context of s. 44 of the Australian Constitution. The user is then given a list of cases, journal article etc., ranked in default by likely order of relevance, to enable them to determine the correct answer to the question (Figure 6).
2. Wherever a citation for a case appears in a dialogue, it will be linked automatically to the text of the case (where it is a neutral citation), with a further link to the LawCite record, as shown in Figure 7.

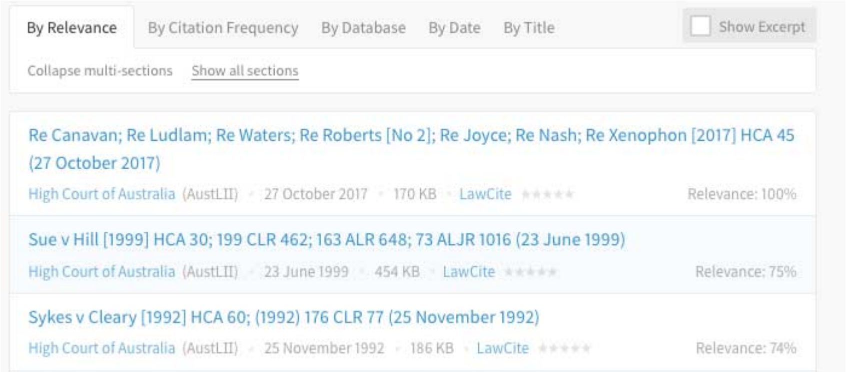


9) Is the nominee under an acknowledgment of allegiance, obedience, or adherence to a foreign power ?

Figure 5. Embedded search link on the words ‘foreign power’

The user is able to note from the LawCite citation record whether that case has been considered by other cases subsequent to the knowledge-base being written, and to check for any resulting changes to the law. No knowledge-base can be updated as frequently as the law might change, and this is particularly so when they are subject to the constraints discussed in Section 2. For example, the LawCite record for this case alerts the user to recent cases considering *Sykes v Cleary*, that may not yet be taken account of in the knowledge-base (Figure 8).

It should be clear from these examples that updating a legal knowledge-base through links and searches requires access to the case and legislation content of a whole legal system, updated continuously. For providers of free legal advice, the most feasible source of such information is a free access Legal Information Institute (LII).



By Relevance | By Citation Frequency | By Database | By Date | By Title | □ Show Excerpt

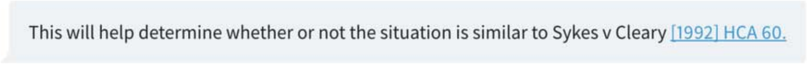
Collapse multi-sections | Show all sections

Re Canavan; Re Ludlam; Re Waters; Re Roberts [No 2]; Re Joyce; Re Nash; Re Xenophon [2017] HCA 45 (27 October 2017)
High Court of Australia (AustLII) · 27 October 2017 · 170 KB · LawCite ★★★★★ Relevance: 100%

Sue v Hill [1999] HCA 30; 199 CLR 462; 163 ALR 648; 73 ALJR 1016 (23 June 1999)
High Court of Australia (AustLII) · 23 June 1999 · 454 KB · LawCite ★★★★★ Relevance: 75%

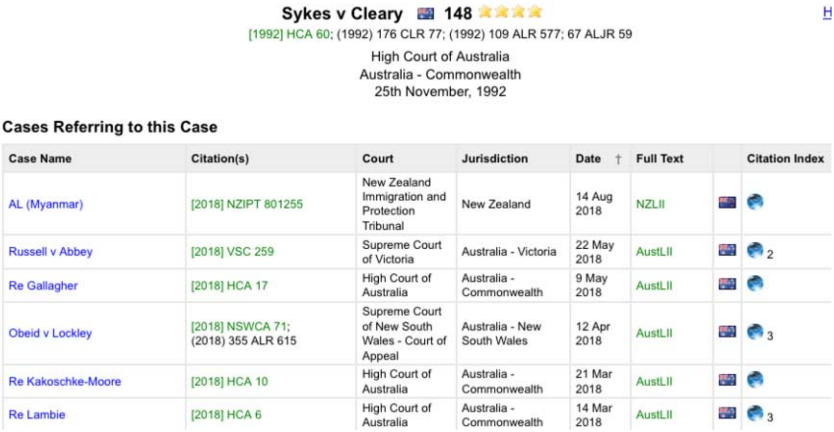
Sykes v Cleary [1992] HCA 60; (1992) 176 CLR 77 (25 November 1992)
High Court of Australia (AustLII) · 25 November 1992 · 186 KB · LawCite ★★★★★ Relevance: 74%

Figure 6. Search results from embedded search link



This will help determine whether or not the situation is similar to Sykes v Cleary [1992] HCA 60.

Figure 7. Automated hypertext links to case law references



Sykes v Cleary 148 ★★★★
[1992] HCA 60; (1992) 176 CLR 77; (1992) 109 ALR 577; 67 ALJR 59
High Court of Australia
Australia - Commonwealth
25th November, 1992

Cases Referring to this Case

Case Name	Citation(s)	Court	Jurisdiction	Date ↑	Full Text		Citation Index
AL (Myanmar)	[2018] NZIPT 801255	New Zealand Immigration and Protection Tribunal	New Zealand	14 Aug 2018	NZLII		🌐
Russell v Abbey	[2018] VSC 259	Supreme Court of Victoria	Australia - Victoria	22 May 2018	AustLII		🌐 2
Re Gallagher	[2018] HCA 17	High Court of Australia	Australia - Commonwealth	9 May 2018	AustLII		🌐
Obeid v Lockley	[2018] NSWCA 71; (2018) 355 ALR 615	Supreme Court of New South Wales - Court of Appeal	Australia - New South Wales	12 Apr 2018	AustLII		🌐 3
Re Kakoschke-Moore	[2018] HCA 10	High Court of Australia	Australia - Commonwealth	21 Mar 2018	AustLII		🌐
Re Lambie	[2018] HCA 6	High Court of Australia	Australia - Commonwealth	14 Mar 2018	AustLII		🌐 3

Figure 8. LawCite records for Sykes v Cleary

4. Ethical and Governance Issues in Free Legal Expertise Systems

In the previous sections we have informally referred to ‘legal advisory systems’ for purposes of readability, but in this Section we need to distinguish between systems which aid interpretation, give advice, and make decisions, so we will used the expression ‘legal expertise systems’ to encompass all three uses, because each of them involves the embodiment and use of expertise.

‘Free legal expertise systems’ are systems which are able to be used for consultation purposes at no cost (including with no required disclosure of personal data) by any person, or which are able to be used by legal advisors working for free legal advice providers in order to advise their clients.

Such 'free' systems include, but are broader than, fully 'open' legal advisory systems, which refer to those systems which are not only free to use, but for which both the knowledge-base, and the software required to operate it, are open to anyone to copy, modify and re-use. Between the minimum condition of 'free use' and fully open legal advisory systems there are many intermediate points, all of which we regard as consistent with 'free and open legal advisory systems'.

Multiple parties may be involved in providing such systems, including pro bono knowledge-base developers (from law firms or academia), intermediaries such as a LII, and free advice providers such as community legal centres (CLCs) or others. In our view, these parties need to consider at least the following issues, and to adopt Principles for the ethical development and governance of such systems. These issues apply specifically to the context of free consultation and free advice, rather than the more general commercial context in which AI is employed in law. Principles adopted should, in our view, be consistent with the principles of free access to legal information.

Here, we are only identifying issues which need to be addressed. In a subsequent paper we will propose the principles which should be adopted to address these issues.

(i) Legal expertise systems to aid interpretation, to give advice, and to make decisions must be distinguished. Such distinctions will determine whether warnings must be given that they do not provide legal advice, or who is responsible for any legal advice provided, or the consequences of decisions being made as a result of their use.

(ii) Legal expertise needs to be based transparently on sources of law, whereas this may not be necessary for other uses of AI in relation to law.

(iii) Legal knowledge-bases may need to have transparent attribution of authors and publishers, and of the date of the law they claim to represent. The role of anonymous authorship is very questionable in relation to legal expertise systems.

(iv) Legal knowledge-bases may need to be transparent, and its text available for free access whenever the system is used.

(v) The logic and assumptions of legal expertise systems implementations may need to be transparent, not only the text of the knowledge-base.

(vi) What is needed to ensure that systems claiming to provide free expertise do not charge end users, directly or indirectly?

(vii) Should end-users of legal expertise systems always be able to remain anonymous to the systems in use?

(viii) How can legal knowledge-bases be licensed appropriately for the effective and expanding provision of free legal expertise?

(ix) How can software for legal expertise systems be licensed appropriately so as to expand provision of free legal expertise?

(x) Legal expertise systems will need to observe emerging principles of ethical use of AI, which are becoming numerous [14]; [15]; [16]; [17].

5. Conclusions – When Is AI Feasible for Free Legal Advice Providers?

In this Chapter we have identified why providers of free legal advice are likely to face significant constraints on the resources available to them to develop and maintain AI-based legal advisory systems, and the implications this has for the types of systems they are most likely to use.

We have set out the approach that AustLII, through its DataLex platform, is taking to facilitate the development of such systems, and how the DataLex approach allows implementation of the guidelines for sustainable legal AI that we have proposed. We suggest that similar approaches could be worth considering by other LIIs.

We have identified ten questions about ethical/governance issues which require consideration by those who wish to develop legal expertise systems for the purposes of free legal interpretation assistance and free legal advice provision. These guidelines and Principles will enable development which is sustainable by the organisations likely to be providing such advice, and to which will contribute to an expanding commons of legal expertise embodied in AI-based tools.

References

[1] Greenleaf, G., Chung, P. & Mowbray, A. (2018). *Legal Information Institutes and the Free Access to Law Movement (UPDATE).* Globalex (Hauser Law School, NYU), http://www.nyulawglobal.org/globalex/Legal_Information_Institutes1.html.

[2] Greenleaf, G., Mowbray, A. & Chung, P. (2013). The Meaning of 'Free Access to Legal Information': A Twenty Year Evolution'. *Journal of Open Access to Law, 1*(1), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2158868.

[3] Greenleaf, G. (2017). Review Essay – Technology and the Professions: Utopian and Dystopian Futures. *UNSWLJ, 40*(1), 302-321, https://ssrn.com/abstract=2973244; or (2017) *University of New South Wales Faculty of Law Research Series, 12*.

[4] Schroeder, M. (Ed.) (2016). *Northern Territory Law Handbook.* AustLII Communities, http://austlii.community/foswiki/NTLawHbk/NTLawHandbook.

[5] Mowbray, A. (2008). *LawCite Citator AustLII*, http://www.austlii.edu.au/lawcite/.

[6] Greenleaf, G., Mowbray, A. & Chung, P. (2018). Building Sustainable Free Legal Advisory Systems: Experiences from the History of AI & Law. *Computer Law & Security Review, 34*, 314-326.

[7] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

[8] Greenleaf, G., Mowbray, A. & Chung, P. (2018). The Datalex Project: History and Bibliography. *UNSW Law Research Paper*, 18-4, https://ssrn.com/abstract=3095897; or (2018) *University of New South Wales Faculty of Law Research Series, 4*.

[9] Tyree, A., Greenleaf, G. & Mowbray, A. (1989). Generating Legal Arguments. *Knowledge-Based Systems, 2*(1), 46-51, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2988931.

[10] Mowbray, A. (2019). *ElectKB Knowledge-base*. AustLII, http://austlii.community/foswiki/DataLex/ElectKB.

[11] *AustLII DataLex Development Tools*, http://austlii.community/wiki/DataLex/.

[12] Mowbray, A. (1996-). SINO Free Text Search Engine – Software (Open Source) (Unix/C), http://www.austlii.edu.au/techlib/software/sino/.

[13] Mowbray, A., Chung, P. & Greenleaf, G. (2016). A Free Access, Automated Law Citator with International Scope: The LawCite Project. *European Journal of Law and Technology, 7*(3); (pre-print) https://ssrn.com/abstract=2768104; *University of New South Wales Faculty of Law Research Series, 32*.

[14] Consultative Committee of Convention 108 (2018). *Guidelines on Artificial Intelligence and Data Protection*. Council of Europe.

[15] European Commission (2018). *Ethics Guidelines for Trustworthy AI*, https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai.

[16] *Universal Guidelines for Artificial Intelligence*, Brussels, 23 October 2018, https://thepublicvoice.org/ai-universal-guidelines/.

[17] Human Rights Commissioner (Australia) (2019). *White Paper. Artificial Intelligence: Governance and Leadership*, https://tech.humanrights.gov.au/sites/default/files/2019-02/AHRC_WEF_AI_WhitePaper2019.pdf.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190023

Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web

Arttu OKSANEN [a,b], Minna TAMPER [b], Jouni TUOMINEN [b]
Eetu MÄKELÄ [b], Aki HIETANEN [c] and Eero HYVÖNEN [b]
[a] *Edita Publishing Ltd. (Finland)*
[b] *Semantic Computing Research Group (SeCo), Aalto University (Finland)*
[c] *Ministry of Justice of Finland*

Abstract. Governments publish legislation and case law widely in print and on the Web. Such legal information is provided for human consumption, but the information is usually not available as data for algorithmic analysis and applications to use. However, this would be beneficial in many use cases, such as building more intelligent juridical online services and conducting research into legislation and legal practice. To address these needs, this Chapter presents Semantic Finlex, a national in-use data resource and service for publishing Finnish legislation and related case law as Linked Open Data for legal applications to use. The system transforms and interlinks on a regular basis data from the legacy legal database Finlex of the Ministry of Justice into Linked Open Data, based on the European standards ECLI and ELI. The published data is hosted on the '7-star' Linked Data Finland service and SPARQL endpoint with a variety of related services available that ease data re-use. Rich Internet Applications using SPARQL for data access are presented as application demonstrators of the data service. In addition, this Chapter presents methods and tools under development to automatically annotate legal texts and to anonymize case law documents prior to their publication on the Web. Anonymization is necessary due to issues of data protection and privacy, and annotation is needed for semantic search and interlinking the documents. The automated approaches could significantly speed up the process and minimize costs of publishing legal documents as Linked Open Data.

Keywords. legislation, case law, linked data publishing, automatic anonymization, automatic annotation

1. Introduction

Governments provide publicly available legal information on the Web usually in the form of HTML or PDF documents targeted to human readers. In Finland, for example, legislation and case law are published as HTML documents in the Finlex Data Bank[1], a

[1] http://www.finlex.fi.

publicly available online service since 1997, maintained by the Ministry of Justice [1]. However, Finlex does not provide publicly available machine-readable legal information as open data, on top of which services and analyses can be built by the ministry or third party vendors.

This Chapter presents Semantic Finlex[2], a national Linked Open Data Service for Finnish legislation and case law. The service hosts and publishes a central part of the Finnish legislation along with judgments of the Supreme Court and the Supreme Administrative Court. All of the datasets are automatically updated regularly.

Our work on Semantic Finlex started in 2012, and the first version of the service was published in 2014 [2]. The data included 2,413 consolidated laws, 11,904 judgments of the Supreme Court, and 1,490 judgments of the Supreme Administrative Court. In addition, some 30,000 terms used in 26 different thesauri were harvested for a first draft of a consolidated vocabulary. During this work, some shortcomings of the initial RDF data model became evident as well as the need for using the then emerging new standards for EU level interoperability. The demo dataset also consisted of only one temporal version (2012) of the statutory law and was not updated. These issues have now been resolved in the work reported in this article.

In the following, we first explicate the motivation and use cases for publishing legislation and case law as linked open data. Then the underlying data models and the data conversion process applied in the service are presented, followed by a discussion on enriching the data with semantic and structural annotations. Then, in Section 5, we introduce the Semantic Finlex publishing platform and semantic portal. In Section 6, data analysis and application demonstrators built on top of the service are presented. Finally, in Section 7 we present our ongoing work to automatically anonymize and annotate legal documents.

2. Motivating Use Cases

Many actors and tasks would benefit from access to legislative and judicial content as data:

Information portals. Within the online services provided by different sectors, it is often necessary to refer to various sections of acts and decrees and display these to users. This requires that such sections be referable and readable as online data. For example, various regulations referring to law are published in the fields of construction, defense, and chemical safety.

The media. Since news on fields such as politics and the business world often refer to various sections of statutes, it is sometimes useful to guide readers to the original legal texts. However, this is not possible if the sections in question are not referable or available in data format.

Juridical online services. In Finland, these include services such as Suomen Laki (Finnish Law)[3] by Talentum Oyj and Edilex[4] by Edita Publishing Ltd, which primarily provide juridical information for professionals in law, such as judges and legal counsels, as well as private persons. Maintaining data in current systems is tedious and largely

[2] http://data.finlex.fi.

[3] http://www.suomenlaki.com.

[4] http://www.edilex.fi.

based on manual work, because the data is not available in a form ‘understood’ by computers, but only as documents in PDF, Word, and other formats.

Legislative drafting. When new statutes are drafted in order to complement and supersede previous ones, the drafters have to examine previous statutes in order to evaluate the effects of the changes and avoid discrepancies. However, semantic information on the various versions of and interdependencies between statutes has been available only in text format.

Editing and publishing of legislative texts. Today, legislation-related information is produced in an inconsistent manner, by using various text formats and index term vocabularies to describe information content. If documents were drafted at the production stage in the form of structured data and in accordance with mutually agreed standards, this would facilitate their further processing and linking to other documents, such as materials in Parliament and in publishing systems such as Finlex.

Intelligent services. Legislative information related to problematic juridical situations, such as divorce or estate distribution, is often scattered between various acts, decrees, and legal practice cases. The availability of statutes and legal cases as such is of little help if the reader, such as an ordinary citizen, finds it impossible to piece the issue together. Presenting legislative documents in a form that can be interpreted by a computer, i.e., as semantic data, would enable the development of more intelligent applications, which would in turn enable making law and justice more comprehensible to citizens. For example, legal texts can be automatically linked to other related texts, legal cases, and vocabularies explaining legal terminology.

Research into legislation and legal practice. The enactment of legislation and legal practice are fields of research in which data analysis methods can be used. The topic of such a research might, for instance, be the impact of EU law to national legal practice [3]. However, data analysis methods require that statutes, the connections between them, and case law-based information on their implementation are available in the form of systematically presented data.

Moreover, authorities in Europe strive to improve the semantic interoperability between EU and Member State legal systems, as the methods in use now for storing and displaying legal documents differ among countries. Therefore, the Council of the European Union has invited the introduction of ELI (European Legislation Identifier)[5] and ECLI (European Case Law Identifier)[6] standards that define common identifier and metadata models for legislative and case law documents by applying Linked Data principles.

3. Conversion to Linked Data

This Section presents the datasets, data models, and the conversion process in use in the Semantic Finlex data service.

[5] Council of the European Union (2012). Council Conclusions Inviting the Introduction of the European Legislation Identifier (ELI). *Official Journal of the European Union*, C 325, 3-11.

[6] Council of the European Union (2011). Council Conclusions Inviting the Introduction of the European Case Law Identifier (ECLI) and a Minimum Set of Uniform Metadata for Case Law. *Official Journal of the European Union*, C 127, 1-7.

3.1. Datasets

The Semantic Finlex service currently consists of four different datasets: *Original legislation.* This dataset consists of approximately 49,000 acts and decrees as they originally appeared in the Statutes of Finland, the official publication of Finnish Law. Besides new acts and decrees, the dataset includes amendments and repeals targeted on previously enacted statutes.

Consolidated legislation. Consolidated texts of acts and decrees incorporate their successive amendments. Editorial work has been carried out by a publishing company. Currently, the dataset includes approximately 3,100 statutes.

Judgments of the Supreme Court. This dataset comprises approximately 5,500 precedents published in the Yearbook of the Supreme Court since 1980.

Judgments of the Supreme Administrative Court. This dataset includes roughly 7,500 judgments of the Supreme Administrative Court from 1987 onwards.

All of the datasets are transformed from different legacy XML formats to RDF adapting the ELI and ECLI specifications. New and updated documents are fetched weekly from the Finlex service and converted into RDF.

3.2. Data Models

As law is not constant but changes over time the RDF data model needs to be able to identify the different temporal versions of the law. Secondly, statutory citations refer to both entire statutes and their individual parts. Therefore, we need to identify the different document parts as well as their temporal versions. Moreover, different language versions of the same document may exist and content can be represented in multiple formats that all need to be identified.

The ELI standard applies the well-established conceptual model FRBR (Functional Requirements for Bibliographic Records)[7] to its ontology definition to distinguish between (1) statutes as such (*work*), (2) their different versions (*expression*), and (3) different content formats (*manifestation*). As the ELI implementation guide[8] states that in multilingual environments, such as in Finland where both Finnish and Swedish are official languages, expressions should be used only to model the different language versions, we use the work level to model temporal versions.

We extended the ELI ontology with our own ontology named SFL (Semantic Finlex Legislation) to define separate classes for the different work level entities, i.e., the legislative documents and the document parts as such (*sfl:Statute* and *sfl:SectionOfALaw*), and their temporal versions (*sfl:StatuteVersion* and *sfl:SectionOfALawVersion*). The extended data model is presented on the left side of Figure 1. Different namespaces and prefixes used in the schemas are listed in Table 1. Statutes and their parts are linked to their temporal versions using the property *eli:has_member*. Temporal versions in turn have two language variants in Finnish and Swedish. These language variants are modeled as instances of the *eli:LegalExpression* class and linked to the temporal versions using the property *eli:realizes*. Finally, the different content formats (text and HTML) of

[7] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records*. K.G. Saur Verlag.

[8] ELI Task Force (2015). *ELI: A Technical Implementation Guide*. Publications Office of the European Union.

the language variants are modeled as instances of the *eli:Format* class and linked to the language variant using the property *eli:embodies*.

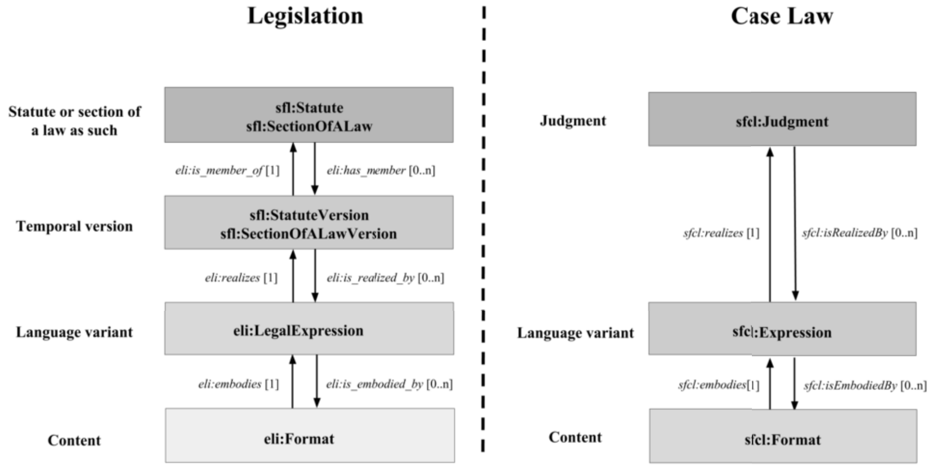


Figure 1. The FRBR-inspired data models for legislative and case law documents

Table 1. Prefixes and namespaces used in the RDF data models

Prefix	Namespace
common	http://data.finlex.fi/common/
dcterms	http://purl.org/dc/terms/
eli	http://data.europa.eu/eli/ontology#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sfcl	http://data.finlex.fi/schema/sfcl/
sfl	http://data.finlex.fi/schema/sfl/
skos	http://www.w3.org/2004/02/skos/core#

SFL further extends the ELI ontology with an additional property *sfl:statuteType* to describe the functionality of a statute, i.e., whether it is a new statute, an amendment, or a repeal. ELI itself defines descriptive properties *eli:type_document*, that we use to describe the level of a statute in the hierarchy of norms (i.e., whether the statute is an act, a decree, or a decision), and *eli:version* to distinguish between original and consolidated document versions.

The physical structure of a statute is modeled at the temporal version level, as it can vary between different temporal versions of the same statute. The properties *eli:has_part* and *eli:is_part_of* are used to model the document tree. Each type of section of a law is modeled as a class in the SFL ontology. The model of the physical structure is presented in Figure 2. As an example, a statute might consist of multiple sections, while sections consist of subsections and subsections consist of paragraphs.

For case law, we adapted the ECLI standard. In contrast to the sophisticated functional model of ELI, ECLI only defines a set of Dublin Core properties to be used to

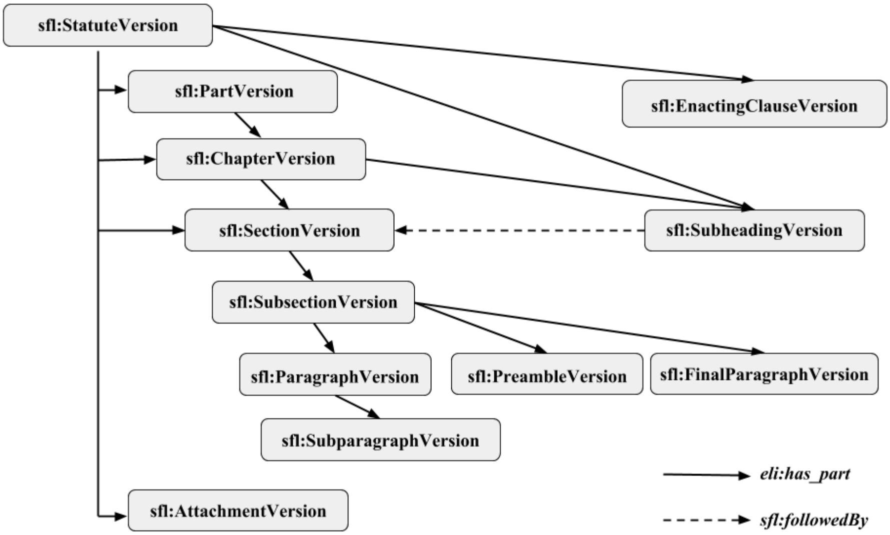


Figure 2. Model of the physical structure of a statute

annotate case law documents. Thus we have developed our own ontology called SFCL (Semantic Finlex Case Law) that wraps the ECLI metadata model to an FRBR-inspired model reminiscent of ELI. However, we can omit temporal versions from the model since there is only one temporal version of each judgment in both Finnish and Swedish. The FRBR model for case law is presented on the right side of Figure 1. Converting the judgments into RDF is quite straightforward as most of the metadata fields mentioned in the definition of ECLI are included in the source data XML.

Contents of both case law and legislative documents are stored at the manifestation level in text and HTML formats as values to RDF properties *sfl:text*, *sfl:html*, *sfcl:text* and *sfcl:html*. This allows the HTML or text content of a specific judgment or section of a law to be retrieved from the triple store with a single SPARQL query.

3.3. URI Identifiers

Besides an ontology, ELI defines a URI pattern schema to unambiguously identify legislation. The URI patterns developed for the Semantic Finlex are presented in Table 2. The original versions are denoted with parameter *alkup* in the URI and consolidated versions with *ajantasa/{date}*, where the date corresponds to the date of entry into force. The documents as well as their parts, their temporal versions, language versions, and content formats can all be identified uniquely using these patterns. As an example, Finnish language version of the Criminal Code of Finland (39/1889) chapter 2 c section 4 as it was in force on February 1, 2018 can be accessed using the URI

http://data.finlex.fi/eli/sd/1889/39/ajantasa/20180201/luku/2c/pykala/4/fin

As for the case law documents, we generate URIs that mimic the ELI pattern, because the standard format for document identifiers defined by ECLI is not an HTTP URI. The URI pattern is presented in Table 3. For example the ECLI identifier

ECLI:FI:KKO:2016:1

is transformed to

http://data.finlex.fi/ecli/kko/2016/1

The document tree structure of the case law documents is not modeled in RDF, and therefore no identifiers for the document parts need to be generated.

Table 2. ELI-compliant URI patterns for legislative documents

URI pattern	Description
/eli/sd/{year}/{id}	Statute as such
/eli/sd/{year}/{id}/pykala/{section id}/...	Section of a law as such
/eli/sd/{year}/{id}/.../alkup	Original (official) version
/eli/sd/{year}/{id}/.../ajantasa/{date}	Consolidated version
/eli/sd/{year}/{id}/.../{version}/{language}	Language variant
/eli/sd/{year}/{id}/.../{version}/{language}/{format}	Content

Table 3. ELI-mimicing URI patterns for case law documents

URI pattern	Description
/ecli/{court}/{year}/{id}	Judgment
/ecli/{court}/{year}/{id}/{language}	Language variant
/ecli/{court}/{year}/{id}/{language}/{format}	Content

3.4. Collecting Version History

To enable access to previous versions of in-force legislation, a version history of consolidated acts and decrees is required. To collect such a history that is queryable by date we need the dates of entry into force of individual statutes and their parts. Resolving the date of entry into force is not always straightforward, since the legacy XML documents do not provide these dates as explicit metadata. Therefore the information must be extracted from the document text using regular expressions which is prone to error.

3.5. Data Validation

Characteristic to the legislative XML formats is the lack of ELI-compliant metadata. In order to produce the RDF metadata it must be extracted from the document text during the conversion process using regular expressions. If the required text is not in the assumed place in the assumed form, then the conversion process may result in missing values and possibly errors in the converted data. Therefore, rules need to be defined to validate the data before allowing it to be published in the service. These rules can be expressed in the form of SPARQL queries performed against the converted RDF data. An example of such a query is one that verifies that the date of entry into force does not precede the date of publication of a statute.

In addition, the ELI validator[9] developed by Sparna Labs can be used to check the conformance of the RDF data against the ELI ontology. The validator applies predefined SHACL shapes[10] (RDF Shapes Constraints Language) to the input data and files a report on the identified constraint violations.

4. Enriching the Data with Relationships and Key Concepts

This Section discusses enriching the data with relationships between different datasets and key concepts describing the contents of the documents.

4.1. Relationships

To describe the interconnectedness of the law, references to different sources of law are extracted from the documents. These relationships are always linked to the most detailed part of the document text.

First of all, the relationship between the consolidated versions and the original amendments is modeled by linking the amendments to the corresponding consolidated versions with the *eli:amends* and *eli:amended_by* properties. This model is depicted in Figure 3. If the section of a law in question has been repealed, a link to the repealing statute is created instead, as presented in Figure 4.

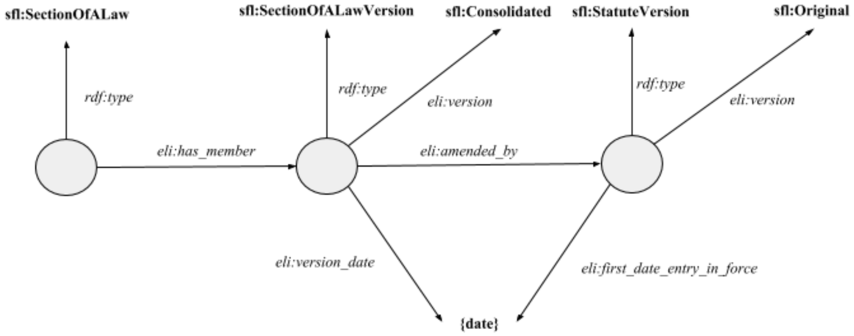


Figure 3. Consolidated versions are linked to the corresponding original amendments.

In addition, references to both national and EU legislation are extracted from the statutes. References to national legislation are denoted using the property *eli:cites*. ELI defines the *eli:transposes* property to describe which legal acts of the EU a statute transposes into national law. The original versions of the statutes contain references to EU directives and regulations for which ELI-compliant URIs can be generated automatically by following the ELI URI pattern.

References to legislative texts are also extracted from case law documents. These are annotated with the property *sfcl:referenceToLegislation*. However, since we do not

[9] http://labs.sparna.fr/eli-validator/.
[10] http://www.w3.org/TR/shacl/.

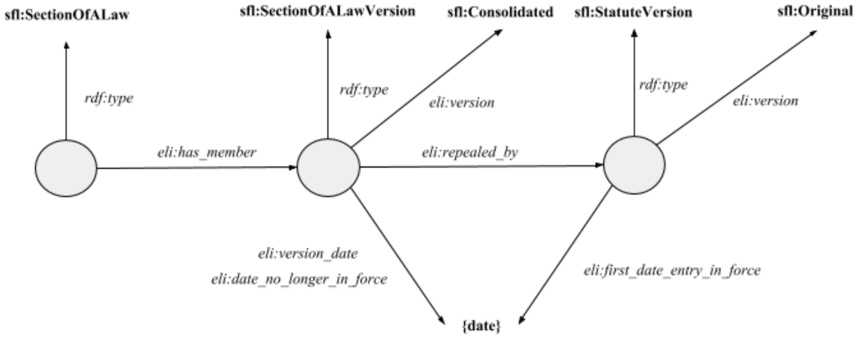


Figure 4. Repealed consolidated versions are linked to the corresponding original repeals.

know which version of a legislative text the citation refers to, we always resolve the link to the abstract work level of a statute or a section of a law, and not any specific temporal version.

To further enrich the metadata of the case law documents, names of the justices of the Supreme Court are extracted from the texts. This is done by using regular expressions that match known types of names. The personnel are modeled as *dcterms:Agent* type of resources and linked to a specific judgment with the property *dcterms:contributor* in accordance with the ECLI specification.

4.2. Key Concepts

To support search and discovery of legal texts, key concepts relating to pieces of legislation were automatically mined from the texts of the documents [4]; [5]. These semantic annotations were selected from the following vocabularies: The Bank of Finnish Terminology in Jurisprudence[11], Eurovoc[12], the legal terminology sections of the KOKO ontology[13], and DBpedia[14].

Before querying the vocabularies, the texts were filtered using stopword lists and linguistic tools. First, the entire text was lemmatized using the SeCo Lexical Analysis Services [6]. The lemmatized results were first filtered based on part-of-speech tags (accepting only words and compound words with proper nouns and nouns) and then the stopword lists were applied to filter out words too general in this context (such as the term *legislation* itself). After this, n-grams from the preprocessed texts were compared against terms in the vocabularies to discover candidate key concepts. Once the results were at hand, the final step was to use a weighting scheme (TF-IDF) to pick only the relevant candidates.

[11] http://tieteentermipankki.fi/wiki/Oikeustiede.
[12] http://eurovoc.europa.eu.
[13] http://finto.fi/koko/fi/.
[14] http://dbpedia.org.

The top scoring[15] candidates were written in RDF format and uploaded to the Semantic Finlex service. The annotations are placed in their own graph[16] using the property *common:autRecSubject* [4]. In addition, the annotations were used to generate tag clouds [7]; [8] to visualize document contents.

5. Publishing Platform and Application Programming Interfaces

The Semantic Finlex service adopts the 5 star deployment scheme suggested by Tim Berners-Lee [9]. The Semantic Computing Research Group has previously proposed an extension to the 5 star scheme by adding two more stars to it [10]. The 6th star is obtained by providing the dataset schemas and documenting them. Semantic Finlex schemas can be downloaded from the service and the data models are documented under the[17] domain. The 7th star is achieved by validating the data against the documented schemas to prevent errors in the published data. Semantic Finlex attempts to obtain the 7th star by applying different means of combing out errors in the data within the data conversion process. The service is powered by the Linked Data Finland[18] publishing platform that along with a variety of different datasets provides tools and services to facilitate publishing and re-using Linked Data.

Following the Linked Data principles all URIs are dereferenceable and support content negotiation by using HTTP 303 redirects. In accordance with the ELI specification, RDF is embedded in the HTML presentations of the legislative documents as RDFa[19] markup. In addition to the converted RDF data, the original XML files are also provided.

To support easier use by programmers without knowledge of SPARQL or RDF, a simplified REST API is provided. This API can be accessed by using the URI patterns and specifying JSON as the preferred content type in the header of the HTTP request. This API also returns its data in the JSON-LD RDF format[20]. Much thought has been given to organize the returned data in a way that is as intuitive as possible and usable also as pure JSON. For example, the affordances provided by JSON-LD *@context* definitions are used to encode language versions of texts in *content_fi* and *content_sv* properties, instead of the user needing to filter the rich literals for their desired language. In addition, URL parameters are provided for retrieving the information pertinent to most common use cases in a stable structure, such as being able to specify which temporal, language, and format versions (of txt and html) of the legislation are required. Finally, a *tree* parameter is provided to build and return the entire subtree of a legislative document without the need to resort to complicated SPARQL queries.

For queries that go beyond fetching information on individual pieces of legislation (such as relational or data analysis queries), a SPARQL endpoint is also available, and a number of sample SPARQL queries are provided to draw inspiration from.

[15] Here, it was decided that the keyword amount is based on document length to have a maximum of 5 to 15 keywords depending on the length of the document. The range was selected based on the analysis of the material [5].

[16] http://data.finlex.fi/annotation/sd.

[17] data.finlex.fi.

[18] http://ldf.fi.

[19] http://www.w3.org/standards/techs/rdfa.

[20] https://json-ld.org/.

6. Application Demonstrators

This Section discusses and presents examples of use cases of the Semantic Finlex data service.

6.1. Data Analysis

Regarding data analysis, sample SPARQL queries were drafted to extract interesting information from the data. These were then fed through to Google Charts for visualization. Information queried was, for example:

- Laws most often referred to from other legislation as well as court decisions.
- Laws that have been changed or amended the most.
- The years in which the above laws were laid.
- The number of EU transpositions by year.
- The members of the supreme court with the most decisions, as well as their tenures.
- The most common topics for supreme court cases by their key concepts.

6.2. Applications

In addition to drafting examples of data analysis, the following application demonstrators were built on top of the Semantic Finlex Linked Data service.

Legal recommender. The HTML representations of the documents are enriched with recommendations to related sources of law similar to [11]. For example, links to relevant EU legislation are queried from the CELLAR SPARQL service[21] by matching their Eurovoc based keywords with the semantic annotations in the Semantic Finlex datasets.

Document-based search. This application can be used to search for similar judgments using an existing document (PDF, image or text) as a search query, for example a court case. The user can choose which search algorithm the application uses to find the documents. Currently the algorithms available are TF-IDF, Doc2Vec and LDA (Latent Dirichlet Analysis).

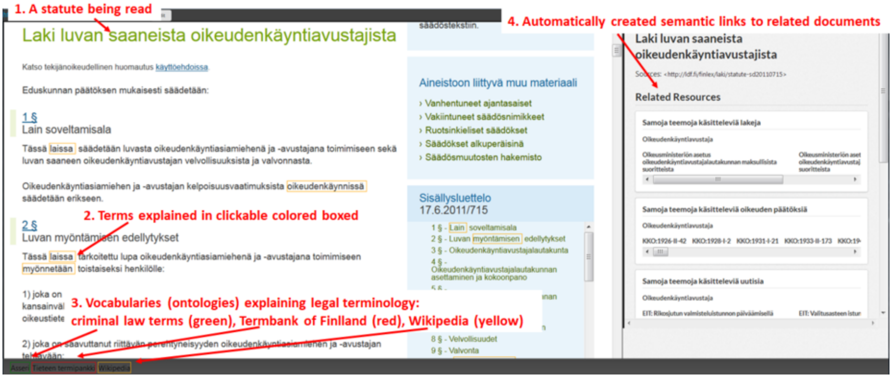
Search based on text and ontologies. Another text-based search tool with semantic autocompletion [12] has been implemented in connection with the service. The search tool works for both legislation and case law. The search is targeted on different fields in the following order of importance: keywords, document titles, section headings, and texts.

Tag clouds. Tag clouds were used to visualize the contents of the documents. For each statute, a tag cloud was generated using the same process as with the semantic annotations of key concepts described earlier.

Contextual reader. To support users in making sense of the legal terminology in the law and court decision texts themselves, the CORE contextual reader [13] was configured with the legal terminology stored in Semantic Finlex. Figure 5 depicts the user interface, where the user is reading a statute (1) in the Semantic Finlex. CORE enables highlighting each instance of specialist terminology in the documents with a popover providing the definition of that term on a mouse-over (2). In this case, three terminological data sources on the Web are connected to the system, indicated by different colors (3). Further, clicking on any marked mention brings in other laws, decisions, and legal news

[21] http://publications.europa.eu/webapi/rdf/sparql.

articles pertaining to that topic, thus facilitating further semantic browsing and delving more deeply into matters either interesting or unclear (4).



1. A statute being read

Laki luvan saaneista oikeudenkäyntiavustajista

Katso tekijänoikeudellinen huomautus käyttöehdoissa.

Eduskunnan päätöksen mukaisesti säädetään:

1 §
Lain soveltamisala

Tässä laissa säädetään luvasta oikeudenkäyntiasiamiehenä ja -avustajana toimimiseen sekä luvan saaneen oikeudenkäyntiavustajan velvollisuuksista ja valvonnasta.

Oikeudenkäyntiasiamiehen ja -avustajan kelpoisuusvaatimuksista oikeudenkäynnissä säädetään erikseen.

2. Terms explained in clickable colored boxed

2 §
Luvan myöntämisen edellytykset

Tässä laissa tarkoitettu lupa oikeudenkäyntiasiamiehenä ja -avustajana toimimiseen myönnetään toistaiseksi henkilölle:

1) joka on kansainväl oikeustiete

3. Vocabularies (ontologies) explaining legal terminology: criminal law terms (green), Termbank of Finlland (red), Wikipedia (yellow)

2) joka on saavuttanut riittävän perehtyneisyyden oikeudenkäyntiasiamiehen ja -avustajan tehtävään;

Asseri Tieteen termipankki Wikipedia

säädöstekstiin.

Aineistoon liittyvä muu materiaali

› Vanhentuneet ajantasaiset
› Vakiintuneet säädösnimikkeet
› Ruotsinkieliset säädökset
› Säädökset alkuperäisinä
› Säädösmuutosten hakemisto

Sisällysluettelo
17.6.2011/715

1 § - Lain soveltamisala
2 § - Luvan myöntämisen edellytykset
3 § - Oikeudenkäyntiavustajalautakunta
4 § - Oikeudenkäyntiavustajalautakunnan asettaminen ja kokoonpano
8 § - Velvollisuudet
9 § - Valvonta

4. Automatically created semantic links to related documents

Laki luvan saaneista oikeudenkäyntiavustajista

Sources: <http://ldf.fi/finlex/laki/statute-sd20110715>

Related Resources

Samoja teemoja käsitteleviä lakeja

Oikeudenkäyntiavustaja

Oikeusministeriön asetus oikeudenkäyntiavustajalautakunnan maksullisista suoritteista

Oikeusministeriön aset oikeudenkäyntiavustaj suoritteista

Samoja teemoja käsitteleviä oikeuden päätöksiä

Oikeudenkäyntiavustaja

KKO:1926-II-42 KKO:1928-I-2 KKO:1931-I-21 KKO:1933-II-173 KKO:19

Samoja teemoja käsitteleviä uutisia

Oikeudenkäyntiavustaja

EIT: Rikosjutun valmisteluistunnon päiväämisellä EIT: Valitusasteen istur

Figure 5. Contextual reader for the Semantic Finlex

Annotation widget. Publishing legislation as Linked Data enables the use of statutes as a reference ontology for linking and integrating heterogeneous datasets that refer to law. The SPARQL endpoint of data.finlex.fi can be utilized as an ontology service [14] for finding relevant statutes or their parts to be used in metadata descriptions. We have prototyped an autocompletion annotation widget that allows the user to search for statutes and fetch their URIs into a cataloging system by typing in a part of the statute name, in the same way as in ONKI [15].

Content widget. As the legislation dataset includes the textual contents of statutes, it can be used to enrich external websites with up-to-date law texts as a web widget [16]. For example, a news article or government announcement informing of a critical change in a statute can be accompanied with the new versions of the relevant parts of the statute. Once the widget is configured with the URI of the desired statute (or its specific part), it will perform a SPARQL query to fetch the text content of the statute, and display it on the web page.

7. Automatic Anonymization and Annotation of Legal Texts

This Section presents our ongoing work in developing tools and services to automatically anonymize and annotate legal texts.

7.1. Automatic Anonymization of Judgments

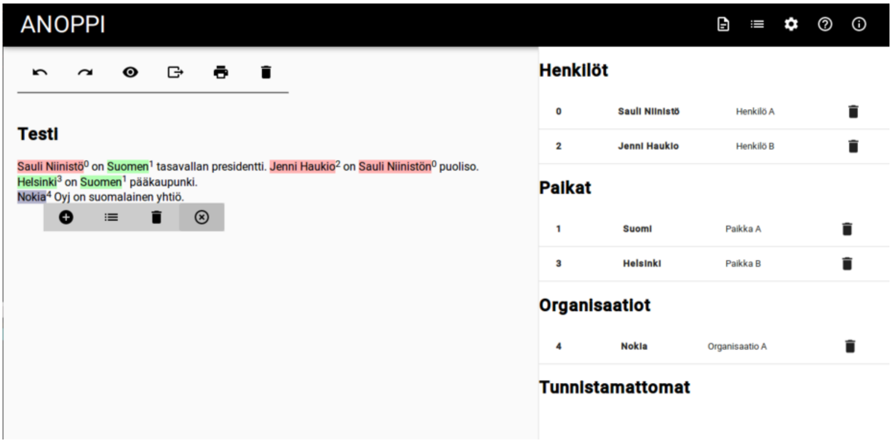
Due to issues of data protection and privacy, judgments must be anonymized prior to their publication as open data on the Web. Anonymization is the process of removing explicitly or implicitly identifying details of persons and organizations from text. In most Finnish courts, anonymized versions of the judgments are not produced during the court process but the set of documents selected for publication on the Web is anonymized manually as needed. However, anonymization is laborous and costly when done manually. Edita Publishing Ltd. has estimated that it takes approximately 40 minutes to manually anonymize a single precedent of the Supreme Court. Yet few automatic anonymization tools are currently being used by the Finnish public sector because of the difficulty of evaluating the adequacy of de-identification for different types of data and requirements [17]. Moreover, it is difficult to find a tool able to carry out the anonymization reliably for both Finnish and Swedish language texts.

To facilitate the publication of court cases on the Web and reduce the costs of anonymizing them, we have started developing a configurable general-purpose anonymization service for Finnish and Swedish texts. The service carries out the anonymization process automatically by performing pseudonymization, that is replacing all identifiable information appearing in a text with neutral identifiers. To take into account the different requirements of different actors we make the service configurable so its users can choose what types of identifiable information need to be obfuscated. Taking into account the possibility of error in the automatic anonymization process, the service allows the user to make revisions on the anonymized version using a web-based editor before exporting the final version.

The anonymization service consists of two separate software components, namely a web service and a user interface. The web service takes text as input, finds the named entities in the text and produces as output the same text annotated with special tags that mark the occurrences of the named entities. The special tags contain additional metadata about the occurrences, such as an entity-specific identifier, a category (person, place, organization etc.), and grammatical case. The identifier and category are required so that the occurrences can be correctly transformed into their correct pseudonymized forms such as ‘person A’ or ‘company B’. The information about the grammatical case is needed so that the occurrences of named entities in the text can be replaced with their pseudonyms in correctly inflected forms. To locate the different types of named entities in the text, natural language processing tools such as part-of-speech taggers and different named entity recognition tools are utilized. The recall of the tools is more crucial than the precision [18], as it is more important to hide all critical information than to have some non-critical information incorrectly hidden.

The user interface, shown in Figure 6 is a web-based editor that allows the user to perform the whole anonymization workflow for a document. The document can be either in plain text, XML or HTML format. The anonymization workflow starts by importing a document into the editor. The document is first processed by the web service finding a set of named entity candidates according to a list of preconfigured categories. The document text with the occurrences of named entities highlighted is then shown in the left column of the application view and a list of the named entity candidates along with their pseudonyms is shown in the right column. The user is then able to modify the named entity candidates and the occurrences. After editing is finished, an anonymized version of the document can be exported with all of the occurrences of selected named entities substituted with pseudonyms.

The performance of the service will be compared with the estimates of Edita Publishing Ltd. The success of the tool depends on the applicability of the user interface in addition to the precision and recall of the language technology tools. Therefore, we will incorporate the users and carry out usability tests already in an early phase of the project.



Testi

Sauli Niinistö^0 on Suomen^1 tasavallan presidentti. Jenni Haukio^2 on Sauli Niinistön^0 puoliso.
Helsinki^3 on Suomen^1 pääkaupunki.
Nokia^4 Oyj on suomalainen yhtiö.

Henkilöt

0	Sauli Niinistö	Henkilö A
2	Jenni Haukio	Henkilö B

Paikat

1	Suomi	Paikka A
3	Helsinki	Paikka B

Organisaatiot

4	Nokia	Organisaatio A

Tunnistamattomat

Figure 6. User interface of the anonymization service

7.2. *Automatic Annotation of Legal Texts*

The same approach used in the automation of the anonymization process can also be applied to the task of automatic or semi-automatic annotation of texts. The difference is that instead of obfuscating the occurrences of named entities found in the text they are linked to existing knowledge bases by applying Linked Data standards. In the context of legislation and case law these entities could be, for instance, key concepts, legal citations, references to parliamentary works or names of judges. Therefore, we plan to use the same user interface and natural language processing tools for the annotation task as for the anonymization, but instead of generating pseudonyms for the named entities, a list of links is provided from which the user can select the correct one. The automatically linked annotations can then serve, for instance, as a basis for the contextual reader application.

8. Related Work and Discussion

Similar efforts to publish legislation and case law as Linked Open Data have been conducted in various countries. The main inspiration for our work was the MetaLex Document Server[22] [19], that provides regularly updated Dutch legislation as Linked Open Data utilizing the CEN MetaLex XML and ontology standards. Another known example of a MetaLex based legal Linked Data service is legislation.gov.uk[23] that hosts UK legislation in local XML formats together with RDF metadata based on the MetaLex ontology. There is also an implementation of a legal Linked Data service in Greece, named Nomothesia[24] [20], that uses both MetaLex and ELI ontologies and ELI-compliant URIs.

Various other ELI implementations and prototypes have also been implemented, usually by resolving ELI-compliant URIs and rendering ELI metadata to existing legal in-

[22] http://doc.metalex.eu.

[23] http://legislation.gov.uk.

[24] http://legislation.di.uoa.gr.

formation portals such as in Luxembourg [25], France [26], and Norway [27]. Many countries already produce ECLI-compliant case law documents to be indexed by the ECLI search engine[28]. Semantic Finlex aims to widen focus by providing both legislation and case law as Linked Open Data through simple Linked Data APIs and linking both of the datasets with each other.

In addition, examples of automatic anonymization and annotation methods applied to legal texts [18]; [21]; [22] and in other domains [23]; [24]; [25] already exist. We aim to apply similar methods to Finnish and Swedish language texts and offer the end-users means to easily revise the possibly incorrect automatic annotations.

A lesson learned during the Semantic Finlex project was that the way the legislation is currently drafted in Finland prevents publishing up-to-date consolidated versions automatically without the need for manual editorial work. To produce such documents without the need for costly editorial work or error-prone automated named entity recognition techniques, the legislative XML and RDF standards, such as Akoma Ntoso [26], should be applied as early as in the legislative and judicial processes where the documents are drafted.

Another issue is that due to copyrights caused by editorial work carried out by the publishing company, we have had to publish the consolidated legislation under a license that restricts its commercial use. The simplest method to circumvent the copyright issues altogether would be to eliminate the need for consolidation by changing the legislative process so that new versions of complete statutes would be published as official versions instead of amendments comprising individual sections.

The new version of the Semantic Finlex service was released on March 10, 2016, and has been in use since. It has been, for example, used in a public hackathon organized by University of Helsinki Legal Tech Lab[29] in October 2017. The development of the service is carried on by further developing the existing application demostrators and tools for automatic anonymization, annotation and validation of the data.

Acknowledgements

The work is a joint effort between Aalto University, the Ministry of Justice, the Ministry of Finance, Edita Publishing Ltd., Business Finland, and Viestintäalan tutkimussäätiö. We thank Risto Talo, Jari Linhala, and Sari Korhonen of Edita Publishing Ltd. for collaboration. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] Hietanen, A. (2009). Free Access to Legislation in Finland: Principles, Practices and Prospects. In Peruginelli, G. & Ragona, M. (Eds.). *Law Via the Internet. Free Access - Quality of Information - Effectiveness of Rights*. European Press Academic Publishing.

[25] http://legilux.public.lu/editorial/eli.

[26] http://www.eli.fr/en/constructionURI.html.

[27] http://lovdata.no/eli.

[28] https://e-justice.europa.eu/content_ecli_search_engine-430-en.do.

[29] https://www.helsinki.fi/en/networks/legal-tech-lab.

[2] Frosterus, M., Tuominen, J. & Hyvönen, E. (2014). *Facilitating Re-use of Legal Data in Applications. Finnish Law as a Linked Open Data Service*. In Hoekstra, R. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 27th Jurix Conference*. IOS Press, 115-124.

[3] Lindholm, J. & Derlén, M. (2015). Festina lente – Europarättens genomslag i svensk rättspraxis 1995–2015. *Europarättslig tidskrift*, (1), 151-177.

[4] Tamper, M. et al. (2017). AATOS – A Configurable Tool for Automatic Annotation. In Gracia, J., Bond, F., McCrae, J., Buitelaar, P., Chiarcos, C. & Hellmann, S. (Eds.). *Language, Data, and Knowledge*. LDK 2017. Lecture Notes in Artificial Intelligence. Springer, 276-289.

[5] Tamper, M. (2016). *Extraction of Entities and Concepts from Finnish Texts*. Master's Thesis, Aalto University, School of Science, Degree Programme in Computer Science and Engineering.

[6] Mäkelä, E. (2016). LAS: An Integrated Language Analysis Tool for Multiple Languages. *The Journal of Open Source Software, 1*(6), http://dx.doi.org/10.21105/joss.00035.

[7] Kuo, B. Y., Hentrich, T., Good, B. M. & Wilkinson, M. D. (2007). Tag Clouds for Summarizing Web Search Results. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 1203-1204.

[8] Hearst, M. A. & Rosner, D. (2008). Tag Clouds: Data Analysis Tool Or Social Signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 160-160.

[9] Berners-Lee, T. (2006). *Design Issues: Linked Data*, http://www.w3.org/DesignIssues/LinkedData.

[10] Hyvönen, E., Tuominen, J., Alonen, M. & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In *Proceedings of ESWC 2014 Demo and Poster Papers*. Springer.

[11] Winkels, R., Boer, A., Vredebregt, B. & van Someren, A. (2014). Towards a Legal Recommender System. In Hoekstra, R. (Ed.). *Legal Knowledge and Information Systems. Proceedings of the 27th Jurix Conference*. IOS Press, 169-178.

[12] Hyvönen, E. & Mäkelä, E. (2006). Semantic Autocompletion. In *Asian Semantic Web Conference*. Springer, 739-751.

[13] Mäkelä, E., Lindquist, T. & Hyvönen, E. (2016). CORE - A Contextual Reader Based on Linked Data. In *Proceedings of Digital Humanities*, 267-269, http://dh2016.adho.org/abstracts/4.

[14] d'Aquin, M. & Noy, N. F. (2012). Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Journal of Web Semantics, 11*, 96-111.

[15] Tuominen, J., Frosterus, M., Viljanen, K. & Hyvönen, E. (2009). ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies As Services. In *European Semantic Web Conference*. Springer, 768-780.

[16] Mäkelä, E., Viljanen, K., Alm, O., Tuominen, J. et al. (2007). Enabling the Semantic Web with Ready-to-use Web Widgets. In *Proceedings of the 1st International Conference on Industrial Results of Semantic Technologies*, 293. CEUR-WS.org, 56-69.

[17] Bäck, A. & Keränen, J. (2017). Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot, http://urn.fi/URN:ISBN:978-952-243-503-3. Liikenne- ja viestintäministeriön julkaisuja 7/2017.

[18] Povlsen, C., Jongejan, B., Hansen, D. H. & Simonsen, B. K. (2016). Anonymization of Court Orders. In *11th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1-4.

[19] Hoekstra, R. (2011). The MetaLex Document Server. In *International Semantic Web Conference*. Springer, 128-143.

[20] Chalkidis, I., Nikolaou, C., Soursos, P. & Koubarakis, M. (2017). Modeling and Querying Greek Legislation Using Semantic Web Technologies. In *European Semantic Web Conference*. Springer, 591-606.

[21] Lesmo, L., Mazzei, A. & Radicioni, D. P. (2009). Extracting Semantic Annotations from Legal Texts. *Hypertext*, 167-172.

[22] Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. & Soria, C. (2004). Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*. Springer, 593-604.

[23] Szarvas, G., Farkas, R. & Busa-Fekete, R. (2007). State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association, 14*(5), 574-580.

[24] Kleinberg, B. & Mozes, M. (2017). Web-based Text Anonymization with Node.js: Introducing NETANOS (Named entity-based Text Anonymization for Open Science). *The Journal of Open Source Software, 2*(293).

[25] Mendes, P. N., Jakob, M., García-Silva, A. & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 1-8.

[26] Palmirani, M. (2011). Legislative Change Management with Akoma-Ntoso. In Sartor, G., Palmirani, M., Francesconi, E. & Biasiotti, M. A. (Eds.). *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*. Springer.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190024

Lost in the Flood? The Library of the Court of Justice of the European Union and Its Foreseeable Future

Fabio PAPPALARDO [1]
Court of Justice of the European Union

Abstract. If free access to the law and to the case law is often considered one of the ways to develop e-justice, access to information about legal publications is somewhat off the radar. This is very strange if we consider the role of legal doctrine in the process of creating law and in the process of applying (*juris-dicere*) the law. The free online catalogue of the Library of the Court of Justice of the European Union gives access to all its bibliographical records and allows everyone who has Internet access to research EU law and other fields of law effectively. This Chapter analyses the tasks of the Library and the collections in order to find out how this could facilitate access to legal information in certain fields and then examine the changes that the library is facing, due to technological development, to the use of electronic resources and to the more and more stringent constraints on financial and human resources.

Keywords. access to legal writing, library of the Court of Justice of the EU, discovery tool

1. Introduction

The Library of the Court of Justice of the European Union is facing a legal data deluge. With the Open data and Open Science movements, with the development of a new means of publishing and sharing legal information through eBooks, eMagazines, Facebook, Twitter and other social networks, it is possible to find a case-note on the very same day as the delivery of a judgment, or to read exchanges between highly qualified scholars in the commentary added to a post on Facebook. Should a library take into account developments of this type? And if so, how should it be done? And how can the quality of the information be guaranteed, when one is shifting from a virtually closed system, a catalogue with chosen records, to an open system that may give access to resources that do not meet the usual quality standards applied to items in our catalogue?

The legal data deluge imposes a new idea of a law library [1], given the capacity to select what should be available in a library, and to reflect also on how the development of

[1] The information and views set out in this Chapter are those of the Author and do not reflect the official opinion of the Court of Justice of the European Union.

law (case law, legislation and doctrine) via the Internet will build the law libraries of the future, and how it will impact the catalogue of the Library. In this short study, we shall briefly recall the mission and the evolution of the Library of the Court of Justice during its 67 years of existence. Next, we shall explain why the Library of the Court of justice has a role that is quite different from most other libraries and how its Internet catalogue compares to EUR-Lex[2] for EU legislation and case law, or to the search engine for the case law that you can find on Curia, the web site of the Court of Justice[3].

And, lastly, we shall examine some new ideas for improving our services, and some alternative approaches to what a jurisdictional library like ours should offer its internal or external users in order to help them to find the information they need.

2. Mission and Services of the Library

Like every other comparable Court, the ECJ possesses a library. Ours was created in 1953, a short while after the Court of Justice of the European Coal and Steel Community was established as a legal body.

The primary task of the Library is to provide assistance to its users, namely the members of the Court of Justice and the General Court and their officials, in carrying out their duties. The main goal of the Library, which is to be a depository for all the documents on European integration [2] is complemented by the promotion of knowledge concerning the European legal system through our online catalogue [3]. As the former registrar of the ECJ pointed out, the Library is an essential working tool for the ECJ [4].

All the documents in the collection are analysed, and a bibliographic record is created for every document that might be of interest to the institution. There is, therefore, a specific record for all the documents concerning EU law or an EU law-related subject [5]. Consequently, not only non-serial publications, but also single articles in periodicals, Libri amicorum and edited collections relevant to the work of the Court of Justice are catalogued and indexed. In a sense, the catalogue, containing such rich subject descriptions, can be compared to a kind of bibliographic database without abstracts. It is updated every day and it is freely available via the Web[4].

The Library's policy of development of the collection depends on the development of the European Union. Every new accession treaty augments the collection by a new national law compartment, and the competencies of the EU law, as modified by the subsequent treaties, must be reflected in the collection. As a consequence of both new accessions and new policies of the EU, the collection now comprises two thousand and forty hundred volumes and two thousand and eight hundred periodicals, together with electronic databases. We are keen to focus on the evolution of the EU, for example, by endeavouring to maintain the different editions of the handbooks on EU law.

The Library purchases between approximately four thousand and six hundred new documents of various types per month.

Over the years, the importance of collecting primary paper version sources has decreased, since reliable and often free-of-charge official publications are available.

[2] https://eur-lex.europa.eu/homepage.html?locale=en.

[3] https://curia.europa.eu/jcms/jcms/j_6/en/.

[4] http://www.bib-curia.eu.

The profile of the collection depends upon the mission of the Library. Even though its aim is to gather a complete collection of material on the European Union legal system, this kind of publication constitutes only 20% of the collection. The remaining 80% includes documents on the national law of the 28 Member States and non-EU countries, general works on comparative law, the theory of law, international law, the history of European integration, politics, economy, and other sources relevant to the staff of the European Court of Justice. Apart from printed documents, there is also a digital collection only available, for the moment, through the internal web site.

The Internet catalogue interface is simple and intuitive, in a certain way Google-like, and available in both French and English language versions. Its records use the MARC21 format with Anglo-American Cataloguing Rules 2 modified according to local needs.

At the moment we have more than half a million records in three different alphabets, sixty languages and thirteen different formats. Even if the aim is to respect the multilingual approach which is intrinsic to the European Union values, a quarter of the records are in German and more than 428,000 records, the 85% of the records concerns the 6 languages used as pivot language by the institution (German, English, French, Italian, Spanish and Polish). If you add to that Dutch, language of 2 founding Member States of the European Communities and the European Union, we reach a total of more than 91% records in 7 of the official languages of the UE.

Subject searching can be carried out using classification codes. The Library worked out its own classification system which better meets its needs than general classification schemes like Dewey or others. The classification scheme comprises almost thirteen thousand five hundred codes including one thousand four hundred terms relating to EU law. A second intranet-only catalogue also allows searching by keywords. Since the internal working language of the Court of Justice is French, these keywords exist only in French. We hope that we shall be able to translate them into all the official languages and to add them to the Internet catalogue; but to do so, we should have to transform the existing system of free keywords created when needed in the structured system of thesaurus.

3. The Legal Academic Writing and the ECJ and Its Library

Historically the focus of the open access movement was on primary sources, law and case law. Attention to Open Science has emerged more recently. Through its catalogue the Library ensures transparency, openness and access to the same bibliographic record on EU law to all citizens. Direct access to the documents has not yet been enabled, since this can be provided only by the authors or by bodies authorized to re-use those documents by the authors, but it is a good start.

I shall try to illustrate the importance of legal academic writings or doctrine by using three examples taken from three different fields of law: comparative law, international law and, of course, EU law.

As was pointed out by one of the most renowned scholars of comparative law, Professor Rodolfo Sacco, statutes, case law and legal writings are the main ‘legal formants’, components that together became the rule applied in particular legal systems [6]. It is unrealistic to compare only the statutes, to analyse only the case law or to study only the different legal writings if we want to find out how a legal system deals with a specific topic. Also, he stressed that, historically, on some occasions doctrinal propositions have been regarded as supreme sources of law.

For this reason, in order to compare the rules applied in different legal systems we must take into account all the ‘formants’, including legal writings. To do so, it is crucial to have access to information on the doctrine.

Very briefly on the subject of International law, I shall just recall that the statute of the International Court of Justice explicitly states that “The Court, whose function is to decide in accordance with international law such disputes as are submitted to it, shall apply … the teaching of the most highly qualified publicists of the various nations, as subsidiary means for the determination of rules of law” (Statute of the International Court of Justice, Article 38).

As far as EU law is concerned, the doctrine exerts its influence on the jurisdictional power in various ways and in many forms. To refer to the most evident form, it is sufficient to point out that direct quotation of a text of doctrine is relatively frequent in an explicit form in the Opinions of the Advocates General, whereas the Court and the General Court in their decisions generally refer to the ‘doctrine’ without giving further details.

Advocate General La Pergola, in the opinion given in the very famous Centros case[5] cited seven different articles or books, six of them concerning the EU law. In our catalogue, there was a bibliographic record for all of them. The seventh was an article on the Abuse of Law published in 1933 in the United States. We have the journal in our collection, but there was no specific record on that article.

It seems useful to stress that the quality of citations has also improved over the years, at least as far as opinions are concerned, with the indications of author, title, journal if needed, year of publication and pages included. Nowadays online sources are also cited. The judgments, on the contrary, which generally have merely an unspecified reference to the doctrine, allow no possibility to check the quotations [7].

This impacts the ability of anyone interested in the case law to find out the basis of the legal reasoning of the Courts on a specific point.

4. EU Legal Scholars and the Library of the ECJ

The catalogue of the library has been online since 2012, but it is not yet very well known. The catalogues of the Commission and of the European University Institute are probably more frequently used than that of the ECJ, but, as far as legal writings on the EU are concerned, the catalogue of the Court of Justice is the most exhaustive and the most complete. To demonstrate this point I shall cite an email exchange between an author and the library.

The first mail was sent at the end of August 2018:

> “Dear Madam/Sir,
>
> I am writing to kindly ask you to add the following bibliographic reference in your web site library databases:
>
> [reference to an article]. Please note that I already addressed this request in person last June and I was told that the Journal volume, where my article was published, was not yet available in the library. Thank you in advance for your availability”.

The Library answered:

[5] ECLI:EU:C:1998:380.

> "In the meantime we have found out that issue [reference to the issue where the article was published] was never received by our Library, in spite of several claims we sent to our supplier. We will do everything we can to obtain the missing issue. Thank you for your understanding".

The author immediately replied:

> "Thank you so much for your prompt reply. Would it be possible to add the reference while waiting for the issue to be delivered? I imagine you probably have an electronic version of the [missing journal] for the internal users of the library.
>
> Thank you again for your availability".

The Library replied:

> "As the measure you propose has several drawbacks, I would kindly ask you to wait until the print copy of the relevant issue will be delivered to us and your article will be indexed and catalogued in the regular way.
>
> As I wrote to you yesterday, we will do all that lies within our power to make sure the missing issue will be sent to us as fast as possible.
>
> I count on your understanding and thank you for the interest that you show in our Library".

The author then replied:

> "Thank you again for your prompt reply. I understand that you catalogue what is available in your library, however, as you know, EU legal scholars rely on your library databases for their bibliographic research. This means that articles published in your missing journal issues will not be included in their work".

Even if the missing article was added to the catalogue when we finally received the journal, a few comments should be made.

This flattering email exchange shows not only that the scientific community understands the value of the Library's catalogue, but also that sometimes it is difficult to explain the difference between a library catalogue and a bibliographic database.

I should also like to recall that prof. Tizzano, former Vice-President of the Court of Justice, proudly mentioned, in his farewell speech given in October 2018 when leaving the Court of justice after 18 years of service, that his "thesis, written in 1963 is still available at the library of the Court of Justice".

The catalogue of the ECJ library grants internal and external users access to the most complete collections on European integration. But from the exclusively paper collections, we are moving to electronic collections. These two types of collections are complementary to each other, with some of the documents available only in one of the two formats. Other types of scientific literature which are having a great impact, and that go beyond the ordinary bounds of scientific publications are case notes and articles published on blogs and on legal websites. In a few years we have moved from paper catalogues, which filled entire rooms, to electronic catalogues and other new options of searching, such as a discovery tool.

For this reason we must reconsider the function of the library in the light of the transformations underway in the scientific publishing system and in the world of access to information provided by libraries. We should also think about the new services that we may offer. Our new discovery tool will be launched in the first semester of 2019 and will ultimately replace the external catalogue.

As Sebastiano Faro and Ginevra Peruginelli pointed out [8], the core ideas behind discovery tools can be summarised as follows: a large index of all kinds of materials coupled with a simple (Google-like) interface, giving patrons the ability to search across a library's entire accessible contents quickly and easily. In this way these tools provide results in a relevance-ranked, integrated list of hard copy, online, and multimedia content.

If these tools have the advantage of offering easy access to a wide set of documents, the disadvantage is that they cannot search, or provide access to all the documents in all our databases. This could have the consequence that users will prefer to use the documents they can find directly from the discovery tool, thus avoiding extending their research to the databases that are not included and searching for documents that exist only in hard copies.

Moreover, the discovery tools usually give access to a large number of documents, but the pertinence of these documents in the specific field of research is not always as high as it should be. The user finds a lot of results and they are often happy with that, but there is the necessity to refine the search, to do some 'fine tuning' in order to get useful results. This could be done by using the catalogue as a selection of pertinent results and giving specific and higher relevance to the documents that are, at the same time, part of the catalogue and of the discovery tool.

5. What Is Next?

We are considering several ways to manage the flood of data. With online journals that are very often accessible free of charge, with legal blogs, with the Open Science project and the idea that all public funded research should be freely available online, we shall have many more documents available online, and we shall then have to decide if and how we should catalogue the most relevant of them.

We then have to devise a means to improve the metadata. We should add multilingual keywords to our catalogue, and we have recently started to add the European Case Law Identifier (ECLI) as a keyword for the case-note. We are considering the possibility of also adding the ECLI in the case-note already catalogued. It is impossible to add ECLI to every case law that is cited in an article, but probably the legal databases will implement this kind of data in the near future. The creation of a European Legal Doctrine Identifier (ELDI) [9] would be of great interest for us too, but the creation of such an identifier should not limit the freedom of research. For this reason it is very important to give due attention to the independence and the autonomy of the body that should attribute such identifiers.

Another way to cope with the increase of data and documents is to create a network of libraries and research institutes in order to share experience, best practice, bibliographic records and altmetric analysis of the impact of the documents added to our catalogues.

The last main area of development is the services we can supply to the users. We are looking into the creation of a chatbot, to help users handle the catalogue, of an "ask a law-librarian online service", of research guides including case law and doctrine about a specific subject, of a helpdesk with library staff that could perform paralegal work, of the implementation of RSS feeds to help users get updates about a specific subject or author, and of the possibility to add comments on a document through the catalogue.

Last but not least, we should like to perform the first 'users' survey' ever at the Library of the ECJ in order to find out the real needs and expectations of internal and external users.

The challenge of the future will consist in adapting the collection and research systems to the potential offered by new technologies. The library of the Court shall continue to be the home of EU legal knowledge.

If free access to law and to case law is often considered one of the ways to develop e-justice, access to the information about legal publications slips somewhat below the radar. This is very strange if we consider the role of legal doctrine in the process of creating law and in the process of applying (juris-dicere) the law.

The free online catalogue of the Library of the Court of Justice of the European Union gives access to all its bibliographical records and allows everyone who has Internet access to research EU law and other fields of law effectively.

6. Conclusions

To bring into focus the mission of the Library, it is necessary to collect all documents reflecting academic studies made on the European integration process, both in paper and digital versions. In this way the Library of the Court of Justice becomes the only keeper and custodian of the whole legal heritage of a united Europe.

This also implies a new challenge for the Library. Digital humanities, by taking root in the space taken by traditional methods of research, data processing and the handing down of its results, force changes in past systems. Widespread electronic databases influence the functioning of the catalogue. Its role is not only to show what a library possesses within a collection but also to indicate databases and collections available online and absent from the shelves. Despite limitations imposed by copyrights and related rights, as well as the tendency to tighten the conditions of licenses by publishers, the Library of the Court of Justice is expected to keep in mind a universal rule underlying the activity of every library, namely widespread and free access to knowledge.

If in the 1950s, when the European Court was established, it seemed impossible that the user of a library could access all the written collection of a library directly from his office, with an instrument similar to a telephone and a screen[6], as today happens with electronic resources, today we are implementing the possibility of carrying out research through a voice assistant that can provide the results of a research [11]. A modern law library must be able to predict future developments in the world of librarianship and make access to its documents and information available to all users, internal and external in the best possible way. The challenge of the future will consist in adapting the collections and research systems to the potential offered by new technologies. So that the court continues to be “not only the house of justice of the European Union, but also that of EU legal knowledge” [3], it is necessary that the Library is able to ensure its document archiving function, while offering all types of electronic services (e-book, e-journals, databases) available today and knowing how to anticipate and to adapt to the new possibilities offered by the technological development.

[6] See [10]. The Author refers to a dream presented by Prof. R. Fano of MIT at a congress on scientific information.

References

[1] Brooks, T. M., Runge, F. L. & Steenken, B. (2016). The Future of Law Libraries. *Law Faculty Popular Media, 8*, 18-23.

[2] Jochen, S. (1977). Documentation Works and Documentation Projects: The Court of Justice of the European Communities. *International Journal of Law Libraries, 1*, 89-94.

[3] Pappalardo, F. (2014). Il contributo della Direzione della biblioteca, ricerca e documentazione della Corte di giustizia, ovvero "Iura novit curia". In Gattinara, G. & Pappalardo, F. (Eds.). *L'Europa dei diritti*. Donzelli, 96.

[4] Grass, R. (2006). Les ressources humaines de la Cour de Justice des Communautés européennes. *Diritto dell'Unione Europea, 4*, 853-864.

[5] Górska, A. & Pappalardo, F. (2015). Library of the Court of Justice of the European Union. *Przegląd Biblioteczny, 4*(83), 533-545.

[6] Sacco, R. (1991). Legal Formants: A Dynamic Approach to Comparative Law (Installment I of II). *The American Journal of Comparative Law, 39*(1), 1-34, 343-401.

[7] Picod, F. (2009). Doctrine et pouvoir juridictionnel. In Picod, F. (sous la direction de). *Doctrine et droit de l'Union européenne*. Bruylant.

[8] Peruginelli, G. & Faro, S. (2019). Tools of Discovery: Opening Doors to Legal Research. In Peruginelli, G. & Faro, S. (Eds.). *Knowledge of the Law in the Big Data Age*. IOS Press.

[9] van Opjinen, M. (2017). The European Legal Doctrine Identifier – A Missing Link? In Faro, S. & Peruginelli, G. (Eds.). *La dottrina giuridica e la sua diffusione*. Giappichelli, 213-227.

[10] Allen, L. E. (1959). Logic, Law and Dreams. *Faculty Scholarship Series. Paper 4517*, 131-144.

[11] Delestre, N. & Malandain, N. (2017). *Du web des documents au web sémantique*. Bois-Guillaume, 15.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190025

The Rutgers Law Library U.S. Congressional Documents Digitization Collection

John JOERGENSEN
Rutgers Law School (U.S.A.)

Abstract. The motivations and processes developed at Rutgers Law Library for digitizing their print collection of United States Congressional hearings and committee prints, dating from 1967 to 2000 are discussed in this Chapter. Both the technical and collection goals of the project, and the important practical details of how it is being accomplished are described. The main theoretical goal was to show how a large scale digitization project could result in a useable, good quality, and sustainable collection while keeping costs at a scale that many institutions might consider affordable. The collection consists of over 25,000 documents. They are committee hearings and other print material that are generated as part of the U.S. Congress' legislative and oversight roles. Although the materials have been unbound, scanned, and checked for quality by hand, most other processes have been automated to minimize cost. Equipment and other expenses have also been kept to a minimum, but without compromise to overall readability, and archival quality.

Keywords. digitization: cost-effectiveness and practicality, U.S. congressional documents, law library

1. Introduction

The Chapter is a description of the Congressional documents digitization project that we have been conducting at the Rutgers Law School Library. This is one of several digital collections actively developed and maintained by the library. The Congressional Documents collection was our first, and most ambitious, attempt at large scale digitization from print. It will be ongoing for several more years, but the project has matured to the point where it can no longer be called experimental, and is an established part of our operations and our collection. The collection currently contains over 18,000 documents, consisting of over 5.5 million page images.

There are several reasons why we chose to develop this collection, and to do the kind of digitization what we have been doing. The first is practical: space in established libraries is always at a premium, so we needed to dispose of a large number of volumes to make room for new material. However, instead of merely throwing the older material away, we decided to digitize. The U.S. Congressional documents were a good candidate, because no one had digitized these materials already, and as government documents, there are no copyright or other restrictions on their distribution.

At the same time, the documents in this collection are a treasure trove of valuable information on matters the U.S. Congress has considered over the years. We have hearings and committee prints of congressional committees from the late 1960's through 2000. They are the transcripts of testimony and other reports generated as committees of the congress perform their governmental oversight functions, and consider new legislation. Matters from the impeachment of presidents to early considerations of global warming are all contained in these pages.

The technical justification for the project was to try to apply new equipment and techniques to find ways to reduce the cost of digitization. The old conventional wisdom about digitization was that readable and searchable document collections would cost in the range of $0.50/page to produce. The majority of this cost consists of the labor involved in either re-keying the printed material, or in performing detailed proofreading and correction of scanned and OCR'ed text[1]. However, with the advent of cheap, fast, high-quality scanning equipment, accurate OCR software, and inexpensive disk storage, it has become possible to change that equation considerably. In fact, our long-term goal for this project was to bring the cost of production to as close to $0.01/page as possible. If that could be done, then any institution wishing to participate in a digitization project could afford to do it. From a library perspective, having to spend millions of dollars on new shelf space, compared with several thousand on digitization, was an easy financial decision to make.

Since the library held these documents as a part of our role as a U.S. government documents repository, our first step was negotiating with the U.S. Government Publishing Office (GPO) for permission to withdraw our the documents from our depository. It is important to note that since there are no electronic copies of these materials from GPO, this was not a question of substituting the electronic for the print under existing depository rules[2] It was necessary to get permission to withdraw them before digitizing. The GPO was willing to consider the project, but required assurances concerning image standards, metadata, and a usable interface before allowing us to proceed[3]. The GPO requirements being met, we were allowed to proceed.

2. Procedures

The following procedures are what we have settled on as the most efficient and effective for our staff and facilities. It will be apparent that they represent a constant balance between quality and available resources. The guiding principle is not to let the perfect get in the way of the good. At the same time, we try to insure that the good really is good. The process starts with verifying our own catalog entry for the document, preparing the document for scanning, the scanning process, and post-scan processing.

2.1. Catalog Preparation

The first step in our conversion process is to insure that our online catalog remains accurate. So, when a set of books is taken off the shelves for digitization, they go straight

[1] Optical Character Recognition.

[2] See e.g.: http://www.fdlp.gov/collections/collection-maintenance/141-substitution-guidelines.

[3] See e.g. Draft Metadata report: https://www.fdlp.gov/news-and-events/321-metadata-report as an example of the standards that needed to be met.

to cataloging. There, a clerk checks them for an accurate record, including a Library of Congress Card Number (LCCN) (which we use as a key for accessioning in the digital collection), and writes the LCCN on the title page. The cataloger then changes the location code to 'electronic document' and inserts the URL where the document will be accessed. The above-mentioned URL is a reference to a CGI program that takes the document's LCCN number as its parameter. It is of the form: https://njlaw.rutgers.edu/cgi-bin/lib/hearing.cgi?file=[LCCN]. At this stage of the processing, clicking on the URL in the catalog will result in a message informing the patron that the document is waiting to be scanned, and that they should inquire with the library should they need it.

2.2. *Physical Preparation for Scanning*

This a straightforward, if labor-intensive process. In order to scan efficiently at high volumes, a sheet fed duplex scanner is needed. This means, of course, that volumes must be unbound and pages trimmed.

Fortunately, most Congressional hearings and prints are very simply bound: most are folded and stapled with no cover (Figure 1). These are easily disassembled with a heavy-duty staple remover. The separated sections are then trimmed along the spine to get loose sheets using a standard guillotine paper cutter. For larger glued volumes, the covers are removed, the folded sections carefully torn away and then trimmed smooth.

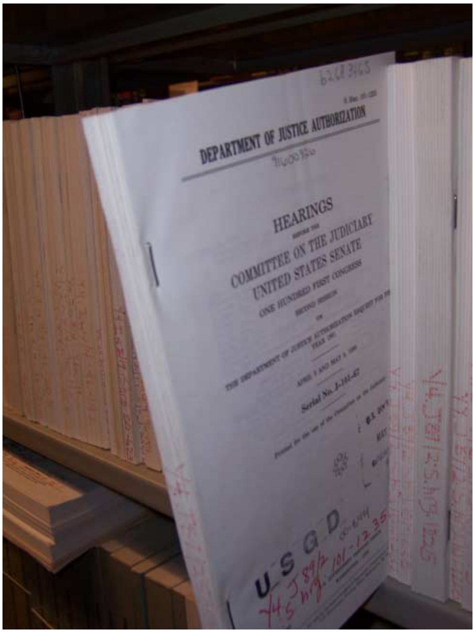

DEPARTMENT OF JUSTICE AUTHORIZATION
HEARINGS
COMMITTEE ON THE JUDICIARY
UNITED STATES SENATE
ONE HUNDRED FIRST CONGRESS
SECOND SESSION
Serial No. J-101-82
U S G D

Figure 1. Congressional document bindings are pretty basic

After some practice, it was found that a little care at the paper cutter helped a great deal. Paper jams and skewed images could be avoided almost entirely by cutting to a consistent width for each volume, as well as being sure to cut enough to insure that all pages were properly separated, but while still leaving a decent amount of margin.

2.3. The Scanning Process

Once cataloged, unbound, and cut, a document is ready for scanning and is placed on a 'ready shelf'. Even with that, we keep standard settings for all documents, so the scan ning clerks only need to load the scanner and enter the LCCN number of the document before hitting the 'go' button. The software creates a new file directory for each new document, named by LCCN. Each scanned image is also named by LCCN, along with a 4 digit sequence number (i.e. LCCN-0001, LCCN-0002, etc.).

2.4. Post Scanning and Quality Control

After scanning, documents are rebanded, and kept until the final processed archival copy of the document has been burned onto DVD. This way they are available for rescanning should any flaws be discovered during quality control.

Once the electronic copy is completely processed, archived and available on the Internet, it is eligible for disposal.

In the meantime, a very basic and quick quality control check is done on images to insure that there are no obvious errors in the scans. This is done using using the Windows Explorer set to 'thumbnail' view (Figure 2). Using this thumbnail view, the checker can easily scroll through the set of images in a document, and identify faulty images. At this stage, defective images are obvious: folded pages, streaks or lines, super- imposed images, etc. The occasional error that is found is rescanned and the defective images replaced. Also verified at this stage are numbering and pagination errors. When all images are verified, a text file containing a copy of the document's library catalog record is placed in the document's directory, and the whole thing is moved to a 'ready' holding directory. The scanned document is now ready for automated processing. The catalog record is in standard MARC (Machine Readable Catalog) format, which is easily reformatted to XML/RDF.

2.5. Automated Processing

As stated earlier, there are two main goals in our preparation of images. First is a quality archive for long-term storage. The other is a readable image for viewing on the Internet. While not exclusive, these are very different goals. Although file size is an issue with the archived image, quality, robustness, metadata availability, and an open standard are more important. Display on the Internet, however, requires as small a file as practically useable, and a format that is easily viewed on most browsers. In addition, the documents should be searchable. Our solution is to create three files, each of which fills one of our goals.

One of the most important aspects of our procedures in this project is that all processing from this point forward is automated. Aside from someone to start the programs running and check on their progress, there is no more labor cost involved in the project. In addition, it should be noted that the processing scrips and programs are, with one exception, produced from open source (as in 'free') software. This also contributes greatly to savings.

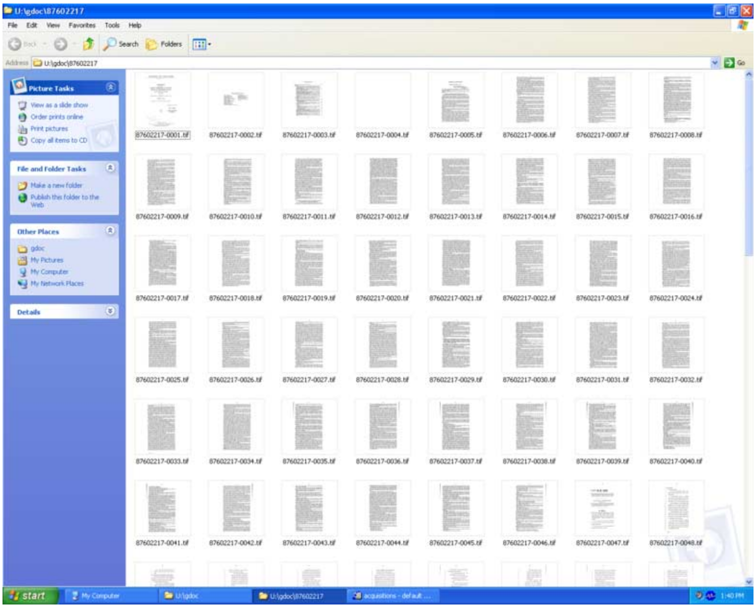


Figure 2. Quality Control: reviewing images for mis-scans

2.6. *Archive File*

One of the most important, and perhaps one of the most overlooked, aspects of successful long-term archiving of computer files is to preserve the connection between a computer file and the metadata which identifies that file. Typically, a digital library creates a metadata repository in the form of a database which has links to the objects identified, the same way a library catalog record has call numbers to direct users to the described book.

Unfortunately, the self-identifying nature of books does not regularly apply to digital records. At best, a whole document contained in a single computer file may or may not be sufficiently self-identifying to allow for some form of cataloging. In the case of something like the Rutgers congressional documents collection, however, each page of a document is stored as an individual set of digital files. If we were to rely on a separate metadata repository to identify items for long term archive purposes, we would be asking for disaster.

Fortunately, built into most current image formats is the facility for reliably embedding information directly into the image file. Although special software is required to write to it, most image formats have registers within the binary code in which descriptive metadata can be inserted. At Rutgers, we use Exiftool to embed bibliographic and other information directly into each image we will be using as an archival copy. This is free software, and can be used in batch-mode to edit many files at a time.

As noted earlier, during preparation of a document, we save a file containing a raw MARC record, taken from our catalog. During processing, this record is read, and a subset of information (author, title, LCCN#, SuDoc#, LC Subject Headings, etc.) are retrieved and reformatted as Dublin Core tags in RDF format. To this is added information on the image itself, such as dimensions, color depth, dots per inch, and where the image fits in the sequence of pages. An example of such a record is at Appendix I.

By embedding this information directly into each one of the images in the collection, each page of each document has sufficient identifying information so that the entire collection can be reassembled even without any external metadata repository. And, it can be done relatively simply. So, even if all filenames were changed and the database we use for searching were to be deleted, the entire library of documents can be reassembled by the computer using the metadata stored in each image file. As long as the image files themselves are maintained intact, we have good assurance of longevity for this collection.

The result is that our archived images are in a well supported open format that will be viewable for forseeable generations of computers. The holographic nature of the embedded metadata provides easily accessible and permanently available information which provides sufficient context to reliably place each image in its place within each document in the collection, no matter what may happen in the future to the external OPAC or interface database that are in place. They will never be just a jumble of images.

2.7. The PDF and OCR'ed Text Files

At the same time as the original scanned TIFF image is prepared for archiving, the processing script also creates two extra files. First, the TIFF is copied into a compressed PDF formatted image. Then, the TIFF is scanned by OCR (Optical Character Recognition) software, which generates a searchable plain-text file. The PDF and text file are saved along with the master TIFF image in the archive, but the PDF and text will mainly be used for the collection's user interface.

Of course, the PDF file is embedded with the same metadata as the TIFF image, also using Exiftool. In the case of the PDF, XMP format, which is particular to PDFs, is used.

As to the OCR'ed text: at this point, OCR software has become surprisingly reliable, 99% accurate in recognizing letters and words in images. With such accuracy, the OCR text can be searched with good reliability even without proofreading. Accurate for searching, however, does not necessarily mean readable. 99% accurate still means that out of an average 1,500 characters on a double-spaced page, there are still 15 errors. In addition, failures in font rendering and layout produce an inconsistent and unsatisfying result. For this reason, the compressed PDF image produced for presentation to the end user. Since it is a clear image, the original formatting, and letter renderings can be interpreted by the user themselves. This way, the OCR text can be indexed by our full-text search engine, providing accurate searching, but the user will view a readable, and quickly transmitted image file in PDF format.

It would, of course, be preferable to have a perfectly accurate and well formatted text file to work with. However, the cost of manual proofreading and reformatting makes this impossible. Given the good searchability of uncorrected OCR, and the viewability of both the PDF and TIFF image files, our method is an acceptable compromise. Complete proofreading, moreover, can only give marginal improvement to search accuracy, but with very large added cost. This must be compared to the presentation of an image, which, if of good quality, will always be a completely accurate representation of the original page.

This is, in fact, the process used in the JSTOR project, and Hein-on-Line, (not to mention Google Books), both of which have proven the viability of these methods[4]. It is not perfect by any means. However, at a generous estimate, we are able to digitize and make available to the world very large amounts of significant material at a cost of something near $0.02/page.

The software used to handle all this processing is, with the single exception of the OCR software, is free open source software that is widely available.

2.7.1. A Short Digression on PDF

The PDF format is very flexible as to content, and can be essentially whatever you need it to be. In the current project, the PDF files in question are pure, unsearchable page images, preserved in the PDF format so they can be conveniently viewed in a web browser. A PDF file can also contain pure formatted text, which would be searchable, or even text with an image superimposed (which is also searchable). In the context of the current project, a pdf/text file would display all the imperfections of the OCR result, and would not be desirable. A pdf image-on-text file would preserve display and be searchable, but suffers from the disadvantage of being a very large file. In the case of documents that are fairly small (say, 25 pages or less), this would be a very workable solution, but with book size documents, it is not practical. Our workstations cannot produce them, our Internet connection would break down transmitting them, and all but the most powerful machines would freeze on attempting to display them.

3. Preparation of a Searchable Document

At this point in the processing of each document, we need to address the problem of context in full text searching. The problem is that individual pages are too small a sample of the typical congressional hearing or committee print to preserve the context of a document as a whole. Searching an index of individual pages is not much use. Since there is no guaranteed regularity to the documents that would allow for any sort of sampling or breakdown, it was decided that the documents must be made searchable only as whole documents. To do this, the OCR'ed text files of each page are concatenated with embedded page break markers inserted at the beginning of each page to reconstitute a single text file of the whole document. The metadata set is then also included as HTML meta tags, as well as some basic HTML head and body tags. The Swish-e search engine then indexes each document as a whole, and creates secondary indicies for each of the meta tags included in the document headers. In this way, users can not only search the full text, but can include some useful fields in their query, such as title, author, year of publication, LC subject heading, and Sudoc number.

Finally, the processing script also concatenates the compressed PDF image files into a set of multipage files of about 50 pages/file. The user interface makes these files available to users who wish to download the enitre document.

In the archive, the compressed PDFs, page level text, large HTML, and large PDF files are saved along with the TIFF master image. All of this material is saved in a single directory per document, and burned onto DVD's as well as being stored on our remote

[4] https://about.jstor.org/ and https://heinonline.org/HeinDocs/HOLBrochure.pdf.

backup, both locally and to a commercial cloud service. The final steps in the processing are that the OPAC is updated to include a link to the online document, and a notation is included in the OPAC and in a separate collection database of the DVD serial number on which the document is archived.

4. User Interface

The core of the user interface for all of the Rutgers – Camden digital collections is the Swish-e search engine[5]. This is a highly configurable, and robust freeware search engine, which remains under active development. It supports boolean searching, can return KWIC (key word in context) sample results, and can maintain separate indicies of html meta tags.

As concerns our government documents project, the configurability of Swish-e is of vital importance. As mentioned above, Swish-e is actually indexing large quasi-HTML files made of contatenated OCR'ed text files. This allows for needed context for searching. When displaying results, however, it would be preferable to display the book page which actually contains the search results. This is something that had to be hacked into the operation of Swish-e. We did it in two steps. First, As mentioned, at the time the large HTML file is created, numbered sequential page break markers are also inserted. Then, we modified the way Swish-e handles its KWIC results feature.

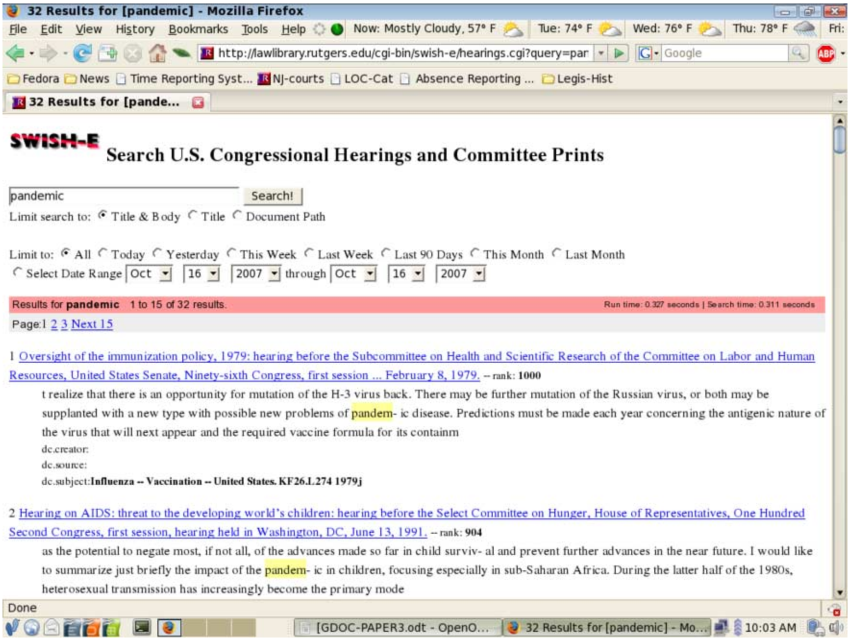
By default, Swish-e provides about 20 or so words before and after the first search term located, enough for most people who wish to look at a KWIC result. We started by expanding this to a much larger number, large enough so that we were sure to capture the numbered page break marker that would come ahead of the keyword. We then added code to the Swish-e script to scan the KWIC result for the page break number that is closest to, but preceding the keyword. We then inserted more code that, instead of providing a link to the indexed HTML document, calls another web script that will present the PDF image that corresponds to the page number, along with a side bar of navigation tools that allow the user to jump between individual page images quickly, or to download the entire document in one or more large 50 page files. That job done, we cut the KWIC result back down to a reasonable size for use on the Swish-e search results page (Figure 3).

Finally, the documents are displayed as seen in Figure 4. Individual page PDF Images are displayed for ease of reading and transmission, along with a navigation panel which allows for paging, etc. A link for downloading the entire document is also provided.

5. Cost and Practicality Issues

Those who might be familiar with producing digital text will note the greatly reduced amount of labor needed to produce a searchable document collection using the above methods. In fact, the labor savings are all the more dramatic when one considers the amount of multitasking that is possible using these methods. Given reasonably careful document preparation, we have experienced very few problems with paper jams, etc. in the scanning process. This means that the scanning devices, once loaded with 100 or so

[5] http://www.swish-e.org.


32 Results for [pandemic] - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Now: Mostly Cloudy, 57° F Tue: 74° F Wed: 76° F Thu: 78° F Fri:
http://lawlibrary.rutgers.edu/cgi-bin/swish-e/hearings.cgi?query=par
Google
Fedora News Time Reporting Syst... NJ-courts LOC-Cat Absence Reporting ... Legis-Hist
32 Results for [pande...
SWISH-E
Done
[GDOC-PAPER3.odt - OpenO...
32 Results for [pandemic] - Mo...
10:03 AM

Search U.S. Congressional Hearings and Committee Prints

pandemic Search!

Limit search to: ● Title & Body ○ Title ○ Document Path

Limit to: ● All ○ Today ○ Yesterday ○ This Week ○ Last Week ○ Last 90 Days ○ This Month ○ Last Month
○ Select Date Range Oct 16 2007 through Oct 16 2007

Results for **pandemic** 1 to 15 of 32 results. Run time: 0.327 seconds | Search time: 0.311 seconds

Page:1 2 3 Next 15

1 Oversight of the immunization policy, 1979: hearing before the Subcommittee on Health and Scientific Research of the Committee on Labor and Human Resources, United States Senate, Ninety-sixth Congress, first session ... February 8, 1979. -- rank: **1000**

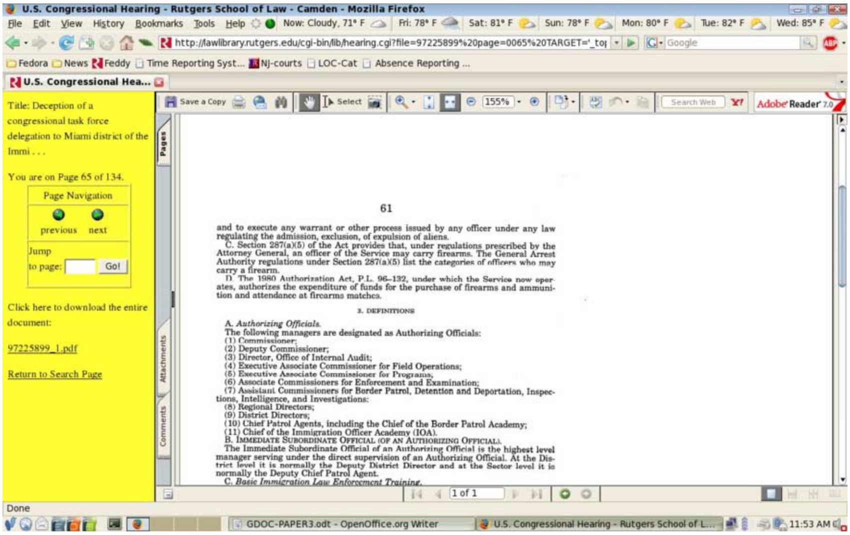
t realize that there is an opportunity for mutation of the H-3 virus back. There may be further mutation of the Russian virus, or both may be supplanted with a new type with possible new problems of pandem- ic disease. Predictions must be made each year concerning the antigenic nature of the virus that will next appear and the required vaccine formula for its containm

dc.creator:
dc.source:
dc.subject:**Influenza -- Vaccination -- United States. KF26.L274 1979j**

2 Hearing on AIDS: threat to the developing world's children: hearing before the Select Committee on Hunger, House of Representatives, One Hundred Second Congress, first session, hearing held in Washington, DC, June 13, 1991. -- rank: **904**

as the potential to negate most, if not all, of the advances made so far in child surviv- al and prevent further advances in the near future. I would like to summarize just briefly the impact of the pandem- ic in children, focusing especially in sub-Saharan Africa. During the latter half of the 1980s, heterosexual transmission has increasingly become the primary mode

Figure 3. Swish-e search results page


U.S. Congressional Hearing - Rutgers School of Law - Camden - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Now: Cloudy, 71° F Fri: 78° F Sat: 81° F Sun: 78° F Mon: 80° F Tue: 82° F Wed: 85° F
http://lawlibrary.rutgers.edu/cgi-bin/lib/hearing.cgi?file=97225899%20page=0065%20TARGET='_to|
Google
Fedora News Feddy Time Reporting Syst... NJ-courts LOC-Cat Absence Reporting ...
U.S. Congressional Hea...
Save a Copy
Select
155%
Search Web
Adobe Reader 7.0
Pages
Attachments
Comments
1 of 1
Done
GDOC-PAPER3.odt - OpenOffice.org Writer
U.S. Congressional Hearing - Rutgers School of L...
11:53 AM

Title: Deception of a congressional task force delegation to Miami district of the Immi . . .

You are on Page 65 of 134.

Page Navigation
previous next
Jump to page: Go!

Click here to download the entire document:

97225899_1.pdf

Return to Search Page

61

and to execute any warrant or other process issued by any officer under any law regulating the admission, exclusion, of expulsion of aliens.

C. Section 287(a)(5) of the Act provides that, under regulations prescribed by the Attorney General, an officer of the Service may carry firearms. The General Arrest Authority regulations under Section 287(a)(5) list the categories of officers who may carry a firearm.

D. The 1980 Authorization Act, P.L. 96–132, under which the Service now operates, authorizes the expenditure of funds for the purchase of firearms and ammunition and attendance at firearms matches.

3. DEFINITIONS

A. *Authorizing Officials.*
The following managers are designated as Authorizing Officials:
(1) Commissioner;
(2) Deputy Commissioner;
(3) Director, Office of Internal Audit;
(4) Executive Associate Commissioner for Field Operations;
(5) Executive Associate Commissioner for Programs;
(6) Associate Commissioners for Enforcement and Examination;
(7) Assistant Commissioners for Border Patrol, Detention and Deportation, Inspections, Intelligence, and Investigations:
(8) Regional Directors;
(9) District Directors;
(10) Chief Patrol Agents, including the Chief of the Border Patrol Academy;
(11) Chief of the Immigration Officer Academy (IOA).
B. IMMEDIATE SUBORDINATE OFFICIAL (OF AN AUTHORIZING OFFICIAL).
The Immediate Subordinate Official of an Authorizing Official is the highest level manager serving under the direct supervision of an Authorizing Official. At the District level it is normally the Deputy District Director and at the Sector level it is normally the Deputy Chief Patrol Agent.
C. *Basic Immigration Law Enforcement Training.*

Figure 4. View of a document. Compressed PDF image displayed with navigation controls to the left

pages at a time, run with very little attention. So, little, that a single staff member typically runs both our scanners simultaneously, while either preparing new books, or performing quality control from a third computer. The staff only pause from their other tasks to load more pages. In the end, therefore, the labor cost per page is actually quite low. The Panasonic duplex scanners have proven themselves to be durable and easy to use, and can be had for under $1,000. The large guillotine-type paper cutter was purchased on E-Bay for about $300.

The other, and even more significant factor controlling the cost is using the ‘dirty OCR’ process. The time and expertise required to proofread the volume of material that we are processing would easily run into the hundreds of thousands of dollars (a rough but reasonable estimate would be something like 800,000 images/year at $0.50/image on average = $400,000). Again, the principle is not letting the perfect prevent the good. The searches are very good, and the images are very readable.

6. Conclusions

Having access to the online information offerings of services like Westlaw and Lexis does not make a library a repository of information for now and the future. It only gives us current access to Thompson and Elsevier’s repositories. To the extent that others may be willing to prepare and actually sell their digital products, we can happily buy and own those items. The strong trend, however, is still in the direction of information being subjected to continuing control by vendors that will continue to charge for access. It seems, therefore, that in order to guarantee public access to the laws of our nations, libraries, LII’s and other organizations need to produce our own digital assets, whether individually or in consortium.

This Chapter is a description of how that can be done in a cost-effective manner, affordable by anyone with the will to start doing it. In the context of a larger consortium effort, the only limit is provided by copyright. Combined with a program of harvesting useful documents from the Internet for preservation and permanent access by the library, a large amount of digital assets that can be accumulated can be significant.

APPENDICES

APPENDIX I: Example RDF Record

```
<?xml version="1.0"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
               xmlns:dc="http://purl.org/dc/elements/1.1/"
               xmlns:dcterms="http://purl.org/dc/terms/"
               xmlns:dctype="http://purl.org/dc/dcmitype/"
               xmlns:dcq="http://dublincore.org/2004/09/20/dcq">
 <rdf:Description rdf:about="URI:/77999007/77999007-0001.tif">
<dc:title>Nuclear order and human values, London, 1977:|report on the eleventh
meeting of members of Congress and of the European Parliament, July 11-13, 1977,
pursuant to H. Res. 313 ... </dc:title>
<dc:creator>
Committe on International Relations, United States House of Representatives
</dc:creator>
<dc:publisher>Washington:.S. Govt. Print. Off.,</dc:publisher>
<dc:source>Y 4.In 8/16:N 88/7 </dc:source>
<dc:contributor>
     <rdf:Bag>
   <rdf:li>United States.|Congress. </rdf:li>
           <rdf:li>United States.|Congress.|House.|Committee on International
```

```
                   Relations. </rdf:li>
           <rdf:li>European Parliament. </rdf:li>
           <rdf:li>Rutgers University School of Law - Camden</rdf:li>
     </rdf:Bag>
</dc:contributor>
<dc:subject>
     <dcterms:LCCN>
           <rdf:value>77999007</rdf:value>
     </dcterms:LCCN>
</dc:subject>
<dc:subject>
     <dcterms:LCSH>
        <rdf:value>
           <rdf:Bag>
              <rdf:li>Atomic power|International control. </rdf:li>
              <rdf:li>Atomic weapons and disarmament. </rdf:li>
              <rdf:li>Civil rights. </rdf:li>
           </rdf:Bag>
         </rdf:value>
     </dcterms:LCSH>
</dc:subject>
<dc:subject>
        <dcterms:LCC>
                <rdf:value>N/A</rdf:value>
        </dcterms:LCC>
</dc:subject>
<dc:language>
<dc:language>
        <dcterms:RFC1766>
                <rdf:value>EN</rdf:value>
        </dcterms:RFC1766>
</dc:language>
<dc:date>
        <dcq:created>
                <rdf:value>1977</rdf:value>
        </dcq:created>
</dc:date>
<dc:date>
        <dcq:issued>
                <rdf:value>2007-1-22</rdf:value>
        </dcq:issued>
</dc:date>
<dc:format>
        <dcterms:IMT>
                <rdf:value>image/tiff</rdf:value>
        </dcterms:IMT>
</dc:format>
<dc:format>
        <dcterms:extent>
                <rdf:value>64284 bytes</rdf:value>
        </dcterms:extent>
<dc:format>
        CompressionType=Group4
</dc:format>
<dc:format>
        ImageSize=3307 x 5423
```

```
</dc:format>
<dc:format>
        ImageDensity=600x600
</dc:format>
<dc:format>
        Colors=Bilevel
</dc:format>
<dc:format>
        ImageDepth=1 bits
</dc:format>
<dc:rights>
        Public Domain
</dc:rights>
<dc:relation>
        <dcq:isPartOf>
                <rdf:value>page 1 of 98</rdf:value>
        </dcq:isPartOf>
</dc:relation>
</rdf:Description>
</rdf:RDF>
```

APPENDIX II: Sample MARC record, source of RDF

```
001       NJRL06-B7891
008       061114s19777777dcu7777777777f000707eng7d
040     __ |a DGPO|DGPO|CStRLIN|NJRL
074     __ |a 1017
086     __ |a Y 4.In 8/16:N 88/7
245     __ |a Nuclear order and human values, London, 1977 :|report on the eleventh
           meeting of members of Congress and of the European Parliament, July 11-13,
           1977, pursuant to H. Res. 313 ...
260     __ |a Washington :|U.S. Govt. Print. Off.,|1977.
300     __ |a xiii, 83 p. ;|23 cm.
500     __ |a At head of title: 95th Congress, 1st session. Committee print.
500     __ |a Submitted to the Committee on International Relations.
500     __ |a Issued Oct. 1977.
650     __ |a Atomic power|International control.
650     __ |a Atomic weapons and disarmament.
650     __ |a Civil rights.
710     __ |a United States.|Congress.
710     __ |a United States.|Congress.|House.|Committee on International Relations.
710     __ |a European Parliament.
852     __ |a NjR-L|Y 4.In 8/16:N 88/7
852     __ |a NjR-L|E-document
856     __ |a |http://lawlibrary.rutgers.edu/cgi-bin/lib/hearing.cgi?
              file=77999007&page=0001|Access this document
902     __ |a ebook
```

APPENDIX III (Personnel Cost Estimates)

Productivity estimates, per 400 page item:	
Unbinding of text:	5 minutes
Scanning (time actually spent on task):	15 minutes
Quality Checking/copying MARC Record:	10 minutes
DVD burning:	1 minute
Processing	0 minutes (done in batches by computer.)
Total estimated time per book	36 minutes/item.
Personnel cost:	$18.00/hour * 0.52 hours/item = $9.30item.
	400 pages/item = $0.023/page.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190026

EU Judicial Procedures and Case Law Databases: What's Going on and What May Lay Ahead?

Elena Alina ONTANU [a,b] and Marco VELICOGNA [b]
[a] *Erasmus University, Rotterdam (The Netherlands)*
[b] *Istituto di Ricerca sui Sistemi Giudiziari – IRSIG - CNR (Italy)*

Abstract. The raise of computational power, the boost of electronic data storage capabilities, and the growing ubiquitousness of the Internet facilitate the collection of legal information and increases its availability for stakeholders. In this context, EU institutions and key stakeholders are seeking to support initiatives that provide access to legislation and case law. This is considered paramount for economic activities, facilitating access to justice, and upholding the rule of law. This Chapter investigates existing electronic databases created to disseminate case law information on the application of EU judicial procedures and explores these databases ability to improve the application of European procedural instruments, forwarding their use and the creation of a common legal understanding. The analysis addresses also the possibilities opened by e-CODEX to integrated cross-national legal database supported by technology developments. The e-CODEX handled cross-border judicial procedures can lead to digital by default judgments in European uniform procedures. These procedures are based on electronic forms supporting structured data exchange. A database relying on these data may be designed to include not only the judgment data, but also many other data generated during the procedure, which could be used to support 'smarter' research for practitioners and interested parties. This can significantly reduce subsequent expert interventions in classifying or anonymizing case law data. Additional data generated by the procedures, but not included in the decisions could also enrich the database. Furthermore, as e-CODEX supports semantic interoperability, much of the structured data is expected to be multilingual by default.

Keywords. case law, e-Codex, electronic procedures, European order for payment, European small claims procedure, legal databases

1. Introduction

Cross-border access to legal information is often a victim of scarce transparency of norms and case law [1]. The raise of computational power, the boost of electronic data storage capabilities and the growing ubiquitousness of the Internet, have facilitated the collection of legal information, increased its availability, and facilitated stakeholders' access to data and documents [2]. Aggregation and dissemination of electronic sentences, court decisions and legal data play a key role in forwarding and sharing knowledge at national as well as at international level. Within the European Union, access to European and national legislation and case law is seen as being of paramount importance for economic activities and for upholding the rule of law. Accordingly, EU institutions and key stakeholders are seeking to support initiatives that go in this direction[1]. Databases and dedicated portals have been created to facilitate access to EU legislation and case law (e.g. EUR-Lex, Curia). Their goal is to support the concrete use of different European legal instruments introduced to facilitate access to justice in cross-border situations. For example, rules relating to European uniform cross-border procedures rely to a significant extent on national frameworks [3]. Accordingly, their implementation and accommodation within these national systems and their functioning in practice are of considerable importance. National case law can play an important role in the way European procedures are applied and interpreted across Member States. Parties and practitioners need to know the specificities of national implementation of EU cross-border judicial procedures and have access to information related to the manner in which uniform rules are applied (in different ways) by national courts. Empirical research has shown that the letter of the EU and national laws is not sufficient [4]; [5].

Although over the last twenty years, access to national court decisions has improved in the European Union and “many courts publish all or at least a substantial selection of their decisions on the internet” [6], accessibility from a cross-border user perspective remains to a certain extent problematic. This is due to several factors, including the lack of common identifiers and metadata, no topical classification, poor formatting of documents, or the provision of scanned documents that are not computer searchable [6].

This is not an issue just from an access to justice perspective. National case law can be inspiring for judges in different Member States who have to deal with the same matter or having similar difficulties in applying a certain provision relating to a European procedure. Additionally, according to the Court of Justice of the European Union (CJEU) CILFIT judgment[2], the national judges have an obligation to consult decisions of other Member States’ courts or of the CJEU, if certain question of European law emerge. This can easily be the case in cross-border litigation involving the application of European uniform procedures or other procedural instruments. The availability of national case law on European procedures can support the creation of a shared understanding of the way uniform rules and procedures should be applied. This can contribute to enhancing mutual trust and sharing of legal knowledge, as judges and practitioners will be able to find out more easily about each other’s work and views ([2], p. 139).

The Chapter begins to consider an important novelty that can represent a new opportunity in developing more integrated cross-national legal databases by making use of the increased availability of electronic procedures at national, and, lately, also at EU level. It investigates existing electronic databases that have been created to disseminate national case law information related to the application of EU judicial procedures. By carrying out this analysis the Chapter looks to determine whether databases are a solution that can lead to the improvement of the application of European procedural instruments, forwarding their use and creation of a common understanding as to their interpretation and application. Furthermore, the Chapter begins to explore the potential implications of using

[1] See for example, European Parliament Resolution of 9 July 2008 on the role of the national judge in the European judicial system, (2007/2027(INI)), O.J. C 294 E/3.12.2009, para 10-11.

[2] CJEU, *CILFIT v Ministero della Sanità*, C-283/81, ECLI:EU:C:1982:335, para. 16. See also ([6], p. 1).

electronic decision resulting from digital procedures. To do this it analyses key features of the European e-Justice Digital Service Infrastructure (e-CODEX), which has been developed to interlink existing national and European ICT systems in the e-Justice domain, and to allow electronic communication and exchange of case related data in cross-border legal procedures.

2. Methodology

The analysis is based on a multidisciplinary approach that combines theoretical and empirical perspectives into national and cross-border databases developments that shape access to legal information and case law, as well as potential developments that can be supported by the present e-justice pilots. The empirical perspective relies on observations, discussions, and informal interviews with persons involved in the development of national and European databases and their maintenance[3]. The authors have been testing the use of several national and European databases as well as contributing to the creating of such databases through European funded research. In addition, the e-CODEX DSI has been studied though an action research approach[4]. In particular, one of the authors has been actively involved in the development, implementation, maintenance, and long-term sustainability effort of the e-CODEX. This has allowed the researchers to gain access to events discussing these developments, activities related to the design and establishment of such tools, and privileged communications that would not be otherwise accessible for scientific research [9]. This approach provided the researchers with the possibility to "perceive reality from the viewpoint of someone 'inside' the case study rather than external to it" ([9], p. 117) and have a deeper understanding of the on-going developments. By choosing this active research interaction the researchers had the possibility to discuss chosen database design and development solutions, choice and generation of metadata, and other technical decisions. Additionally, this has allowed the researchers to test data interpretations and possible improvements through semi-structured and informal discussions with other researchers and/or companies involved in projects developing legal databases at national and/or European level.

3. Legal Databases: National, Cross-Jurisdiction and European Approaches

Databases are infrastructures that can be designed and built to hold certain type of data and enable specific type of analysis and queries [10]. They 'unmoor data analysis' and, once in place, they enable users to conduct extensive surveys and analysis of data without needing to compile and organise the data themselves [10]. Thus, enabling the spreading of knowledge they contain. In the legal domain, the aggregation and dissemination of legal data is a key point in sharing knowledge, support access to justice, enhance mutual trust, and can contribute to the creation of a common understanding and practice in the

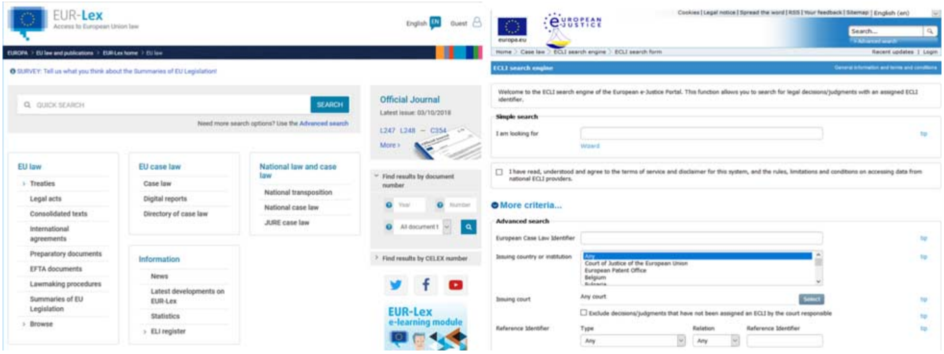
[3] *Setting Up a Case Law Database*, Workshop IC2BE Project (Max Planck Institute Luxembourg for Procedural Law, 26 February 2018).

[4] "Action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework". [7]. See further on action research [8].

application of European instruments. This concerns not just the EU level, as national law plays a significant role in the way the European procedures are applied and function across Member States.

In order to uphold the rule of law and support mutual trust and mutual recognition principles, the European Commission has supported the development of European portal-based access to case law, especially for facilitating access to national cases concerning the application of different European procedural instruments. At national level, different initiatives exist, including private or general national databases that are publicly managed and financed.

The process of setting databases providing access to national and European case law in the area of cross-border litigation and civil cooperation has so far involved different initiatives undertaken by the European institutions or supported by them. The first steps in this direction have included the creation of databases in the area of cross-border litigation accessible through the EUR-Lex and the EU e-Justice Portals (Figure 1). A few initiatives can be singled out: namely, the JURE collection containing national judgments related to jurisdiction, recognition, and enforcement in civil and commercial matters[5], the National case law web page dedicated to EU case law[6], and the e-Justice ECLI search engine[7].



SURVEY: Tell us what you think about the Summaries of EU Legislation!

Need more search options? Use the Advanced search

Welcome to the ECLI search engine of the European e-Justice Portal. This function allows you to search for legal decisions/judgments with an assigned ECLI identifier.

I am looking for:

□ I have read, understood and agree to the terms of service and disclaimer for this system, and the rules, limitations and conditions on accessing data from national ECLI providers.

European Case Law Identifier:

Issuing country or institution: Any

Issuing court: Any court

□ Exclude decisions/judgments that have not been assigned an ECLI by the court responsible

Reference Identifier:

Figure 1. European database initiatives

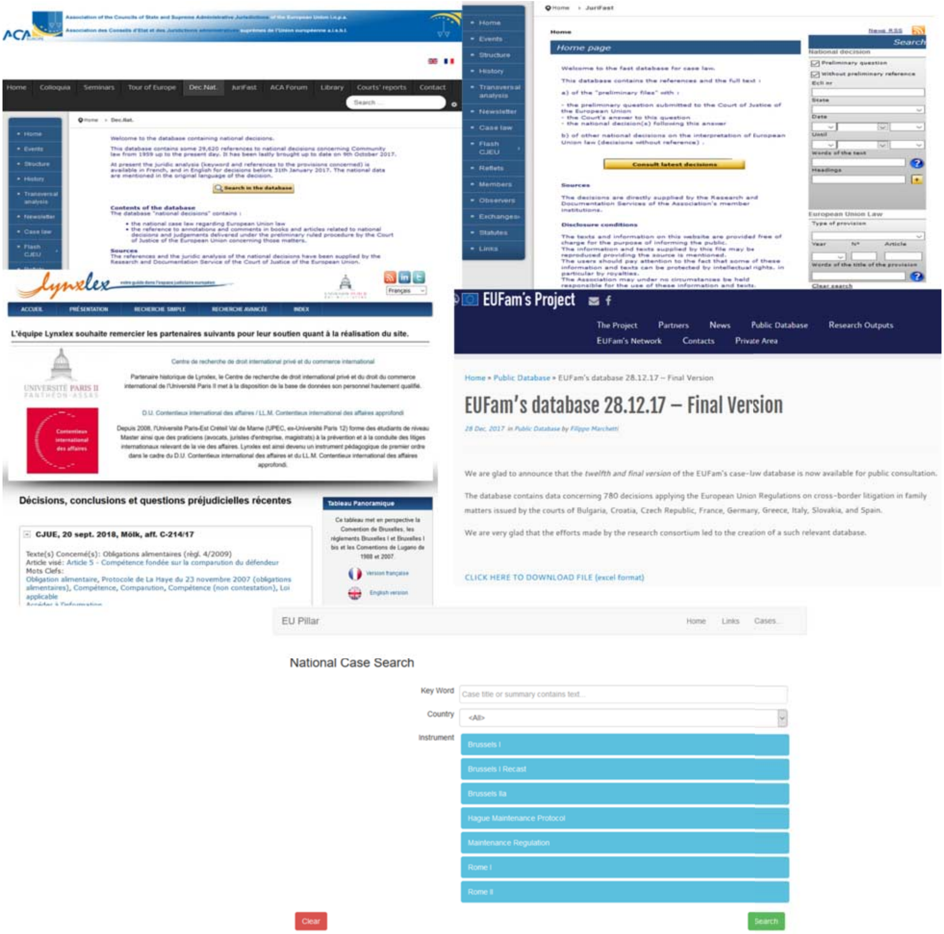
In all these databases, the decisions are generally available in their original language, and sometimes – when provided by the Member State of origin – a summary in English, French, or German (JURE), key words on the topic of the case in English and French (National case law), or, occasionally, an abstract or key words in English or France (e-Justice ECLI search engine). In addition to these European direct endeavours, various initiatives have emerged as the result of work carried out by academics and professional organisations (in some cases supported by EU grants) to develop databases that facilitate

[5] The database was created by the European Commission, and includes case law on relevant international conventions (i.e. 1968 Brussels Convention, 1988 Lugano Convention), http://eur-lex.europa.eu/collection/n-law/jure.html.

[6] http://eur-lex.europa.eu/collection/n-law/n-case-law.html.

[7] The ECLI search engine includes case law from 13 Member States (Belgium, Czech Republic, Germany, Estonia, Greece, Spain, France, Croatia, Italy, Latvia, Netherlands, Slovenia, and Finland). More information on the European Case Law Identifier (ECLI) can be found at https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-en.do?clang=en.

access to national and European case law. These initiatives focus primarily on providing access to case law in relation to the application of European regulations in the area of private international law. In doing so, they strive to maximise dissemination of information by elaborating as much as possible English summaries available to stakeholders in free access. Some examples in this sense are JuriFast and Dec.Nat[8], Lynxlex[9], EUFam[10], and EUPILLAR[11] databases (Figure 2).



Welcome to the database containing national decisions.

This database contains some 29,620 references to national decisions concerning Community law from 1959 up to the present day. It has been lastly brought up to date on 9th October 2017.

At present the juridic analysis (keyword and references to the provisions concerned) is available in French, and in English for decisions before 31th January 2017. The national data are mentioned in the original language of the decision.

Contents of the database
The database "national decisions" contains :

- the national case law regarding European Union law
- the reference to annotations and comments in books and articles related to national decisions and judgements delivered under the preliminary ruled procedure by the Court of Justice of the European Union concerning those matters.

Sources
The references and the juridic analysis of the national decisions have been supplied by the Research and Documentation Service of the Court of Justice of the European Union.

Welcome to the fast database for case law.

This database contains the references and the full text :

a) of the "preliminary files" with :

- the preliminary question submitted to the Court of Justice of the European Union
- the Court's answer to this question
- the national decision(s) following this answer

b) of other national decisions on the interpretation of European Union law (decisions without reference) .

Sources

The decisions are directly supplied by the Research and Documentation Services of the Association's member institutions.

Disclosure conditions

The texts and information on this website are provided free of charge for the purpose of informing the public.
The information and texts supplied by this file may be reproduced providing the source is mentioned.
The users should pay attention to the fact that some of these information and texts can be protected by intellectual rights, in particular by royalties.
The Association may under no circumstances be held responsible for the use of these information and texts.

National decision
☑ Preliminary question
☑ Without preliminary reference
Ecli nr
State
Date
Until
Words of the text
Headings
European Union Law
Type of provision
Year N° Article
Words of the title of the provision
Clear search

Lynxlex
votre guide dans l'espace judiciaire européen
Français
ACCUEIL
PRÉSENTATION
RECHERCHE SIMPLE
RECHERCHE AVANCÉE
INDEX
UNIVERSITÉ PARIS II PANTHÉON-ASSAS
Contentieux international des affaires
Version française
English version

L'équipe Lynxlex souhaite remercier les partenaires suivants pour leur soutien quant à la réalisation du site.

Centre de recherche de droit international privé et du commerce international

Partenaire historique de Lynxlex, le Centre de recherche de droit international privé et du droit du commerce international de l'Université Paris II met à la disposition de la base de données son personnel hautement qualifié.

D.U. Contentieux international des affaires / LL.M. Contentieux international des affaires approfondi

Depuis 2008, l'Université Paris-Est Créteil Val de Marne (UPEC, ex-Université Paris 12) forme des étudiants de niveau Master ainsi que des praticiens (avocats, juristes d'entreprise, magistrats) à la prévention et à la conduite des litiges internationaux relevant de la vie des affaires. Lynxlex est ainsi devenu un instrument pédagogique de premier ordre dans le cadre du D.U. Contentieux international des affaires et du LL.M. Contentieux international des affaires approfondi.

Décisions, conclusions et questions préjudicielles récentes

CJUE, 20 sept. 2018, Mölk, aff. C-214/17

Texte(s) Concerné(s): Obligations alimentaires (régl. 4/2009)
Article visé: Article 5 - Compétence fondée sur la comparution du défendeur
Mots Clefs:
Obligation alimentaire, Protocole de La Haye du 23 novembre 2007 (obligations alimentaires), Compétence, Comparution, Compétence (non contestation), Loi applicable
Accéder à l'information

Tableau Panoramique

Ce tableau met en perspective la Convention de Bruxelles, les règlements Bruxelles I et Bruxelles I bis et les Conventions de Lugano de 1988 et 2007.

Home » Public Database » EUFam's database 28.12.17 – Final Version

EUFam's database 28.12.17 – Final Version

28 Dec, 2017 in Public Database by Filippo Marchetti

We are glad to announce that the *twelfth and final version* of the EUFam's case-law database is now available for public consultation.

The database contains data concerning 780 decisions applying the European Union Regulations on cross-border litigation in family matters issued by the courts of Bulgaria, Croatia, Czech Republic, France, Germany, Greece, Italy, Slovakia, and Spain.

We are very glad that the efforts made by the research consortium led to the creation of a such relevant database.

CLICK HERE TO DOWNLOAD FILE (excel format)

National Case Search

Key Word Case title or summary contains text...
Country <All>
Instrument
Brussels I
Brussels I Recast
Brussels IIa
Hague Maintenance Protocol
Maintenance Regulation
Rome I
Rome II

Figure 2. Academic and professional organisations developed databases

[8] JuriFast is a database for case law containing 'preliminary files' (i.e. preliminary questions submitted to CJEU, the CJEU decision, and the national decision(s) following the CJEU judgment) as well as national decisions on the interpretation of European Union law, www.aca-europe.eu/index.php/en/jurifast-en). Dec.Nat. is a database containing some 29,620 references to national decisions concerning Community law from 1959 up to 9 October 2017, www.aca-europe.eu/index.php/en/dec-nat-en.

[9] www.lynxlex.com.

[10] www.eufams.unimi.it/.

[11] www.abdn.ac.uk/law/research/eupillar-database-559.php.

All these databases concern specific aspects of EU and national law. The data collection concerns national case law related to the application of EU legislation (e.g. Lynxlex) or cover more than one jurisdiction (e.g. EUFam, EUPILLAR, Bruusels I-bis[12], forthcoming IC2BE[13]). Previous research carried out in the framework of the project 'Building on the European Case Law Identifier (ECLI)' on online publication of court decision in all Member States of the EU revealed that legal provisions and policy frameworks on the publication of case law differ across EU Member States. In practice, not all national courts publish their decisions, provide open access, or including the same level of detail, accessibility, metadata[14], and translations in other languages [2]. Furthermore, the ontologies used differ across databases [11].

The national, cross-jurisdiction, and European projects dedicated to the creation of databases are often the result of labour intensive analysis carried out by experts who select and classify relevant court decisions to produce structured data that is then made publicly available. They all aim to provide structured information that is easy to search through for courts, practitioners, researchers, policymakers, or parties. This can raise awareness on availability of European instruments and support informed choices in cross-border enforcement. Such endeavours can facilitate legal searches and information gathering by users who might otherwise not have been able to access and read these judgments. However, the process is made more difficult by the lack of interoperability and connection between existing databases. As van Opijnen remarks these collections remain quite small while 'the costs for development and technical maintenance are relatively high' ([6] p. 21). Project specific collection of case law are often small and do not manage to attract the attention of the communities of practitioners they are meant to serve as 'many online resources are competing for the lawyer's attention' [6]. Very specific and limited collection of cases are necessary for a small number of cases [6]. Thus, while the ECLI search engine and the European portals integrated databases can be more easy to find for the user, other projects or national initiatives remain to a certain extent less visible for users not familiar with the projects or national initiatives in other Member States as there is no single electronic environment in place providing access to the various existing databases. *De lege ferenda*, a single electronic environment would be a useful European initiative. What can appear at first as small national data has the potential to becoming 'Big Data' by aggregation [12], which in turn can lead to an exponential increase of information offered and made available for further use across the EU.

In general, these database initiatives are characterized by a patchy approach to the selection of the areas they are covering, the timeframe surveyed, their dissemination, the language(s) used, and the manner in which the data are structured. Furthermore, the way data is presented is not usually uniform, for example when access to full case law text

[12] www.asser.nl/brusselsibis/. This database is part of a research project (JUST/2014/JCOO/AG) assessing the application of Regulation (EU) 1215/2012. The database looks to integrate court decisions from the EU Member States that are publicly accessible and apply this regulation (links to the original texts and references to court decisions are not available in free access).

[13] A new research project IC2BE (*Informed Choices in Cross-Border Enforcement*, JUST-JCOO-CIVI-AG-2016) intends to set up an English-language database for national and CJEU cases in relation to the European Enforcement Order (EEO), the European Order for Payment (EOP), the European Small Claims Procedure (ESCP), and the European Account Preservation Order (EAPO), www.ic2be.eu.

[14] This should be understood as a set of specific information selected in relation to the analysed case law included in the created databases.

is provide this is not always computer manageable (i.e. paper document scans). Some of the initiatives create English summaries together with the text of the decision in the original language. The need of translation is added to that of classification and generation of structured data. A complete translation of national cases is too costly and time consuming. Therefore, the process in general focuses in making specific kind of metadata available in English (e.g. type of procedure used, legislative provisions referred to, correlations with other relevant case law and literature, key words) and sometimes summaries. This process is laborious and requires a long process of case selection, data structuring and information choice that has to be carried out in accordance with a previously agreed taxonomy. Furthermore, compliance with data privacy requirements has to be observed[15]. In most situations, the process is purely manual and relies on a legal expert to carry it out. This makes the process highly resource demanding and costly. This has in turn effects for the maintenance, sustainability, and continuity of database initiatives, especially for grant-funded project endeavours.

Approaches that are more inclusive and less resource dependent need to be considered. As legal procedures are increasingly carried out online and judicial decisions are produced electronically, new possibilities are becoming available. Hence, technological solutions may have the potential to support more inclusive and automated data collection process and should be further considered.

4. Technology Input: Main Perspectives on Legal Database Development

Technology input into legal databases development is twofold: namely, through technology that renders access to collection of case law and case law datasets (e.g. Curia, EU-Fam, EUPILLAR), and through direct electronic support and communication via information and communication technology (ICT) with professional bodies and courts. Most of present endeavours rely on technology to provide access to legal databases that have been manually compiled following a laborious and costly process of data selection and metadata creation. The ECLI search engine (ECLI-SE) can be seen as an intermediary step between technology that facilitates access to collections of case law and the full technology based solution. This is because the Member States that implement ECLI and connect their repositories to the ECLI-SE will have their decisions and data automatically made available and displayed via this technical interface ([6], p. 10-11).

Technology based solutions follow a different approach. Technology is increasingly used to set up integrated electronic filing and handling of claims which subsequently leads to a more or less automatic feeding of the electronic decisions, thus, produced in dedicated national and/or European databases. In achieving this, the adoption of a standardised description of the documents is necessary and guarantees interoperability between the national information systems, involved authorities, and used sites. For example, the adoption of XML techniques as standard for representing the documents uni-

[15] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1, 4.5.2016; Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119/89, 4.5.2016.

fies the cross-border referencing system and secures interoperability. Implementing specific XML standards means that 'the users are guided to produce legal documents with a well-defined structure' [13]. This facilitates an automatic detection of specific type of information and/or text regardless of the language in which this is drafted[16].

The availability of structured documents and metadata can lead to a tailored access to information and for specific categories of users. Technology is able to provide and support personalised access to information via established databases. In practice, this would mean that no *ex ante* decision on the content of the text would need to be taken as personalised access to information would deliver specific level of detailed access based on the identity of the user (e.g. judge, enforcement officer, lawyer, party, policymaker). All information and data could be uploaded onto the national and/or European databases and, depending on the type of user consulting the database certain type of information, would become available upon research queries. This process has to comply with legislative requirements (e.g. GDPR and Data Protection Directive) related to processing of personal data, pseudonymisation, consent of data using, privacy issues, and ways in which data can be used and accessed.

Technology based solution that can lead to an integrated electronic registration and handling of cases and their direct upload on dedicated databases offer the opportunity to document the use and application of specific procedures such as the European procedures, to monitor their handling, and measure various aspects of their functioning. This is of significant importance for practitioners in terms of knowledge sharing and for policymakers.

This can be particularly true for European cross-border judicial procedures such as the European Order for Payment (EOP) and the European Small Claims Procedure (ESCP). These European procedures rely on the use of standard forms. This can support the collection of data structured *ab initio* and uniformly across the EU, as well as speeding up its collection and aggregations of decisions and case information that can be subsequently used. This process is eased when the procedures are handled electronically. Section 5 will explore the features of the electronic version of these procedures, which the e-CODEX DSI enables.

5. e-CODEX to Legal Database: New Possibilities for Data Structure and Data Access

An important novelty that can represent a new opportunity in developing more integrated cross-national legal databases results from the possibility of making use of the increased availability of electronic procedures at national, and, lately, also at EU level. This technology and law development may open up and facilitate the collection and availability of national court decisions. These developments result in digital by default judgments, which are the result of electronic procedures. At EU level for example, within the e-CODEX DSI supported judicial procedures, the EOP and the ESCP court decisions can be generated as XML and .pdf documents from electronic filed standard forms and structured data exchanged during the procedure.

e-CODEX DSI has been developed to connect and allow legally valid communication between existing national and European e-Justice back-end systems through a

[16] On the handling of multilingual complexity, see, e.g., [14]; [15].

decentralised network of e-CODEX access points. These access points consist of two components, a Gateway that establishes and ensures secure communication with other e-CODEX access points and a Connector that carries out the adaptations of the messages and their content between national back-end systems standards and the e-CODEX ones (and vice-versa). The generic flow of a message through the e-CODEX e-Delivery solution is sketched in Figure 3.

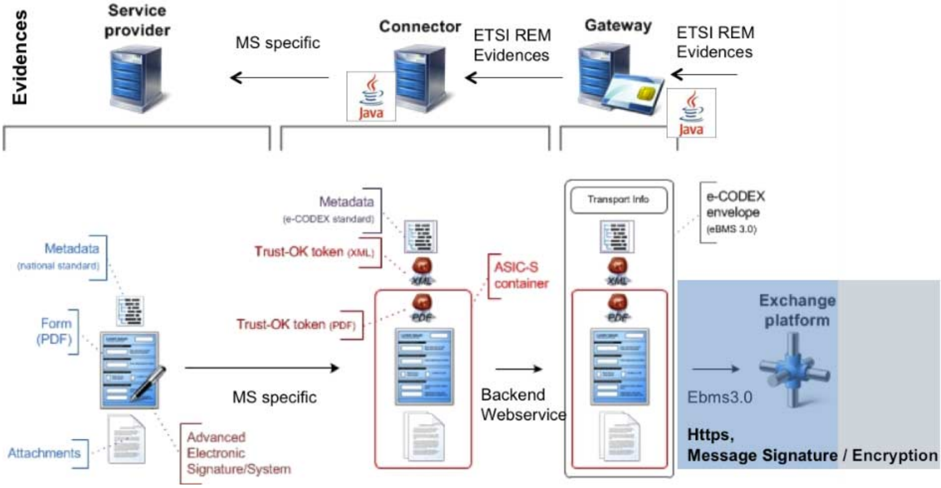
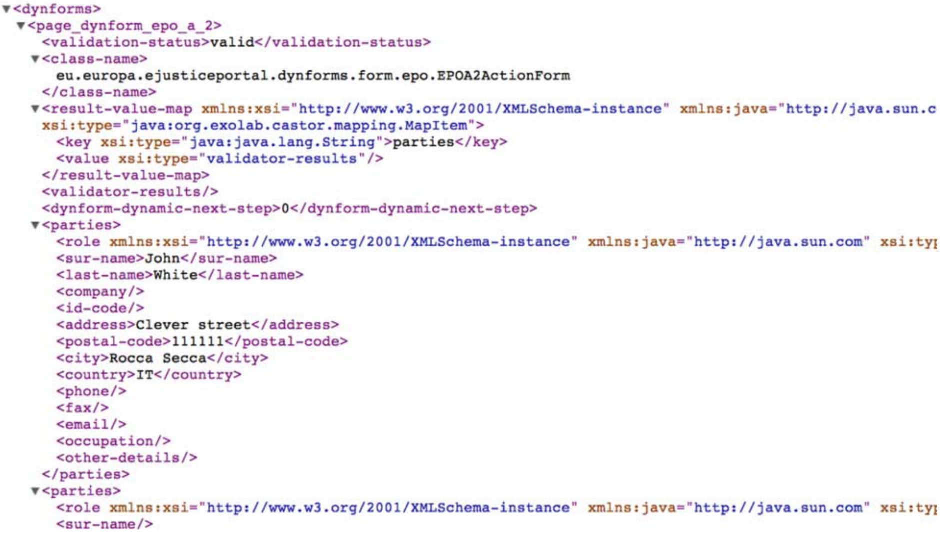


Figure 3. Generic e-CODEX national implementation and outgoing communication flow [16]

The e-CODEX DSI therefore ensure technical and legal, but also semantic interoperability. The semantic interoperability ensures that the correct meaning of the data exchanged is preserved and understood by all systems involved. National back-end e-Justice solutions ‘are typically based on domestic semantic structures. To support the exchange of semantic information, e-CODEX uses common document standards and semantics. Specific coding schemas used by national systems need to be transformed in order to be interpreted by other systems using different schemas. This transformation is better known as mapping’ [17]. ‘Following a use-case centric modelling approach, for each use-case, with the support of national experts, e-CODEX has developed specifications which ensure mutually equal interpretation of data exchanged between national electronic systems in cross border legal procedures’ [17].

At present the European procedures decisions are just sent to the parties and kept by the court. The e-CODEX system is not saving the data that is generated by the proceedings, but changes can be considered in order to use existing data and make use of the system to collect documents and their metadata (that is already unified at EU level) and feed it into a database.

Although the e-CODEX has been developed as a transportation system and it was not envisaged for the creation of an European case law database for the European uniform procedures, the forms based procedures and the metadata generated by the national systems and by the EU e-Justice portal in the transmission and receiving of procedures is a data richness that can be stored in the future and further used (Figure 4 provides an example of XML generated by the e-Justice portal when an EOP form A is filled online). Such data collection (metadata) can allow a more sophisticated and extensive analysis



```
▼<dynforms>
  ▼<page_dynform_epo_a_2>
    <validation-status>valid</validation-status>
   ▼<class-name>
      eu.europa.ejusticeportal.dynforms.form.epo.EPOA2ActionForm
    </class-name>
   ▼<result-value-map xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:java="http://java.sun.c
    xsi:type="java:org.exolab.castor.mapping.MapItem">
      <key xsi:type="java:java.lang.String">parties</key>
      <value xsi:type="validator-results"/>
    </result-value-map>
    <validator-results/>
    <dynform-dynamic-next-step>0</dynform-dynamic-next-step>
   ▼<parties>
      <role xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:java="http://java.sun.com" xsi:typ
      <sur-name>John</sur-name>
      <last-name>White</last-name>
      <company/>
      <id-code/>
      <address>Clever street</address>
      <postal-code>111111</postal-code>
      <city>Rocca Secca</city>
      <country>IT</country>
      <phone/>
      <fax/>
      <email/>
      <occupation/>
      <other-details/>
    </parties>
   ▼<parties>
      <role xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:java="http://java.sun.com" xsi:typ
      <sur-name/>
```

Figure 4. Example of an EOP Form A XML structured data as generated by the e-Justice Portal

of the decisions issued on the basis of the European procedures and of their handling by courts. Such data will be generated by the procedure itself and will not require human intervention.

A database alimented by these electronic decisions could include not just text but also the structured XML data of the decision. This comprises many of the elements of the case that could be used to support ‘smarter’ research for practitioners and interested parties. This may significantly reduce the need of subsequent work related to the classification of data by experts or the task of anonymizing specific fields. Additional data generated during the procedure, but not available in the decision could also be used to enrich the database. Furthermore, as the support of the procedure is multilingual and e-CODEX DSI supports semantic interoperability, all the structured data would be multilingual by default.

In addition, the e-CODEX infrastructure and the richness of the available structured data could be used to profile the access of specific categories of users or to provide for different levels of authorization enabling and managing access to different sets of data, processing requests and authorizations to access data. This could allow the creation of various levels of database research details, the establishment of pre-defined and advanced research criteria and terminology, and provide support for users in need of intuitive assistance. Partial or full anonymisation can be carried out depending on the objective of the search and of the authorisation level of the user (e.g. public prosecutor carrying out an investigation on malpractice in cross-border procedures or a legal professional studying case law).

While a centralised database solutions could be developed, the decentralised nature of e-CODEX DSI could also allow the connection of national databases alimented by e-CODEX procedures. In this perspective, the e-Justice Portal could provide a single entry point through which the queries could be conveyed.

6. Concluding Remarks

This Chapter has provided an overview on the existing landscape of EU legal information databases and investigated a number of possibilities that may be available to improve it. An integrated European database containing extensive information and decisions issued by national and CJEU judges in relation to the application of European procedures and instruments can contribute to the creation of a common understanding and practice in the application of the European legislation. Furthermore, it can facilitate the correlations between various interpretations, characteristics of the cases, and outcome of the procedures. This approach can support judges' and legal practitioners' work, as well as that of parties, and to a certain extent the general public based on the level of access that is retained to comply with personal privacy legislation and GDPR provisions.

In terms of architecture of the system, the e-CODEX DSI connector structure could be used to support an interrelation of central and national databases structures through e-CODEX access points. As the database is envisaged as a direct collection process of electronically handled cases and files, attention has to be paid to the treatment of personal data and legal solutions granting various level of access for different categories of users (e.g. courts, practitioners, parties). Further research should address these legal implications.

In addition to the possibility of supporting access to central and national databases, e-CODEX services generate structured data and documents which could be used to create a decisions database. An e-CODEX database incorporated in the e-Justice Portal and building on the European 'produces' structured data (XML and XSDs) can certainly support practitioners in their interpretation and application of the regulations.

A further integration could be envisaged linking an e-CODEX generated database to the e-Justice ECLI search engine. This further development appears desirable in supporting a 'one stop shop' approach to legal information and case law across the EU. This would enhance cooperation and enable practitioners to become acquainted with the justice system and legislation of other Member States, which will in turn help boost confidence in each other's legal systems, encourage legal professionals to share best practices, and foster mutual trust.

References

[1] Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S. & Soria, C. (2005). Automatic Semantics Extraction in Law Documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*. ACM, 133-140.

[2] van Opijnen, M., Peruginelli, G., Kefali, E. & Palmirani, M. (2017). Online Publication of Court Decisions in Europe. *Legal Information Management, 17*(3), 136-145.

[3] Ontanu, E. A. (2017). *Cross-Border Debt Recovery in the EU. A Comparative and Empirical Study on the Use of the European Uniform Procedures*. Intersentia.

[4] Mellone, M. (2014). Legal Interoperability in Europe: An Assessment of the European Payment Order and the European Small Claims Procedure. In Contini, F. & Lanzara, G. F. (Eds.). *The Circulation of Agency in e-Justice: Interoperability and Infrastructure for European Transborder Judicial Proceedings*. Springer, 254-264.

[5] Ng, G. Y. (2014). Testing Transborder Civil Procedures in Practice: Findings from Simulation Experiments with the European Payment Order and the European Small Claims Procedure. In Contini, F. & Lanzara, G. F. (Eds.). *The Circulation of Agency in e-Justice: Interoperability and Infrastructure for European Transborder Judicial Proceedings*. Springer, 265-286.

[6] van Opijnen, M. (2017). Gaining Momentum. How ECLI Improves Access to Case Law in Europe. *Journal of Open Access to Law, 5*(1), 1.

[7] Rapoport, R. N. (1970). Three Dilemmas of Action Research. *Human Relations, 23*, 499-513.

[8] Reason, P. & Bradbury, H. (Eds.) (2008). *The SAGE Handbook of Action Research*. Sage, 2nd ed.

[9] Yin, R. K. (2014). *Case Study Research: Design and Methods*. Sage, 5th ed., 116-117.

[10] Kitchin, R. (2014). *'Conceptualising Data'. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, 1-26.

[11] Mommers, L. (2010). Ontologies in the Legal Domain. In *Theory and Applications of Ontology: Philosophical Perspectives*. Springer, 265-276, https://openaccess.leidenuniv.nl/bitstream/handle/1887/17578/143317_1_En_12_Chapter_OnlinePDF.pdf?sequence=1.

[12] Lupton, D. (2018). How Do Data Come to Matter? Living and Becoming with Personal Data. *Big Data and Society*, I-II, 1-11.

[13] Francesconi, E. (2015). Legal Knowledge Modeling for Managing Legislation in the Semantic Web. In Araszkiewicz, M. & Płeszka, K. (Eds.). *Logic in the Theory and Practice of Lawmaking*. Springer, 515.

[14] Francesconi, E. (2014). An Interoperability Approach for Enabling Access to e-Justice Systems across Europe. In Kö, A. & Francesconi, E. (Eds.). *Electronic Government and the Information Systems Perspective*, 3rd International Conference, EGOVIS, 26-40.

[15] Smitz, P., Sanmartin, F., Francesconi, E., Hajlaoui, N., Batouche, B. & Stellato, A. (2018). Automatic Alignment of Multilingual Resources in the Linguistic Linked Open Data Cloud. *Journal of Open Access to Law, 6*(1), 1-12.

[16] Velicogna, M. et al. (2014). *D7.4 Architectural Hands on Material*. e-CODEX Project Deliverable, 26.

[17] Velicogna, M. & Steigenga, E. (2016). *Can Complexity Theory Help Understanding Tomorrow E-Justice?* Paper presented at the Conference on Complex Systems, Law and Complexity session (Amsterdam 20-23 September 2016), 13, https://ssrn.com/abstract=2914362.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190027

A Model of Justice as a Platform: A Case Study of Open Data Disclosure

Giulio MICHETTI, Arianna TONIOLO, Simone ROSSI and Alessandro PIRANI [1]
C.O.Gruppo (Italy)

Abstract. Digital technologies influence operations and managerial processes in Courts of Justice. Notwithstanding peculiarities, Courts are incrementing the amount of available data that represent a huge opportunity. Proposing an exploratory case study, this Chapter aims to connect the reflection on new policy making models to the reflection on digital practices in the judiciary offices. An open data disclosure process is a case proposed to reflect about a shift in public administrations' model of interaction with organizational environment: from revising processes to face operational emergencies (extractive interaction model) to revising processes to provide knowledge and interpretative tools to its environment (platform interaction model). Judiciary services produce a platform providing a deep understanding of social and economic systems. Enabling justice to become a platform, where different subjects can acquire information and data, is a goal of primary importance for the elaboration of public policies, but also to empower social innovation and entrepreneurial opportunities.

Keywords. open data, open government, digital transformation

1. Open Data, Open Government and the Role of Public Administration

To perform their tasks, many public administrations produce and collect data. The quantity and sensitivity of data make them particularly significant as a resource for increasing transparency, to provide a better understanding of government actions and increase accountability [1]. Data disclosure could produce several benefits of different types. The phenomenon of and the discussion about open data start from this assumption and can be described in several ways.

Considered as an object, open data are sets of data, generated by the action of a public administration or at its disposal, which are made public and made available – on the Internet – to anyone who wants to consult or elaborate them.

Data can be defined as the lowest level of abstraction from which information and then knowledge are derived [1]. Open data initiatives disclose raw data that can be managed and processed by anyone. This leads to other fundamental characteristics of open data as objects: data must safeguard the privacy of those involved in public proceedings; the sharing format must be manageable without the use of specific non-free software.

[1] The Authors are organizational consultants for C.O.Gruppo (www.cogruppo.it) which is a consulting company specialized in organizational change and digital transformation. It has developed a strong experience in the judiciary system over last 20 years, working with all professionals in the field to develop effective practices and innovative organizational processes.

On the other side, taken as a concept and as a movement of opinion, open data represent an evolution of public administrations' action: data produced and used by government and public administrations should be available to all [2]. In this sense, the debate on open data is strictly linked to the theme of open government, or rather the political ideal that demands citizens' right to the utmost transparency in relation to the public administrations that operate against them.

The literature presents a long list of open government's benefits for societies: a government more transparent, more efficient and the possibility for citizenship to affect decision-making processes [2]; the potential to increase participation, interaction, self-empowerment and social inclusion of open data users and providers [3]; the stimulation of economic growth [4].

Open government debate is fully consistent with the evolution in the way of analyzing and framing the work of the public administration that has occurred in the last few decades which can be defined by three evolutive stages.

The first stage is the highly bureaucratic and hierarchical public administration, close to bureaucracy Weberian ideal type. This model was challenged in the last decades of 20th century by the affirmation of New Public Management (NPM): the public administration should be able to incorporate managerial issues from the private sector increasing control and monitoring of process inputs and outputs; standards of quality and productivity had to be fixed and entrepreneurship within organizations had to be stimulated and rewarded [5].

From several years, even the NPM appears as a superseded paradigm. The new model that has been gradually affirming in the literature, but above all in practice, is New Public Governance (NPG): an idea of plural and pluralistic public action, focused on the management of inter-organizational networks, processes and outcomes of processes, alternative to the institutional structures of private companies [6]. NPG is a paradigm that fully recognizes the fragmentary and uncertain character of public action in the new millennium, where several actors outside the public administration, linked by trust bonds, play a key role in the production and implementation of public action [7].

In the NPG model, the administrative action proposed is that of Public Value Management: a new way of managing processes able to offer greater recognition to a wide audience of stakeholders, to activate deliberative decision-making, open to dialogue with instances of citizenship and negotiation for organizational goals [8]. In this perspective, there is a clear claim for public policy implemented with an active role of stakeholders and citizens.

2. The PCT and the Digitalization of Italian Judiciary System

The empirical case and the elaboration that will be presented in this Chapter spring from a long digital transformation process occurred in Italy in the last decade.

Processo Civile Telematico, or PCT, is the name used to refer to the digital reform occurred in Italy in civil judiciary system. PCT was fully operational a few years ago thanks to the legal obligation of electronic filing for lawyers and professionals, but the change process began more than ten years ago.

Nowadays, nonetheless several pitfalls and opportunities to improve effectiveness, PCT can be considered a successful policy and one of the most relevant public administration's digital transformation – perhaps on the European level – because of the impressive amount of data and official documents stored in and generated by the system. The database on which the system is based collects a huge amount of information on any civil procedure handled by an Italian Court.

More than this, PCT introduced new forms of interactions between Courts and primary procedures actors, and transformed Court offices' work, introducing new practices and operational goals. Together, they introduced a strong datification of civil Court operations.

Datification can be defined as the digital commutation of any activity provided by informatic systems, software and online platforms. The adoption and the utilization of digital devices and softwares are constantly producing data about activities performed that can be exploited to improve work effectiveness and produce knowledge.

3. A Model of Justice as a Platform

In last decades and throughout PCT digital transformation, the relationship between Courts and their organizational environment can be defined by an *extractive model*. This interaction is primarily based on operational needs: Courts engage the environment when acknowledging a shortage of resources of different kind.

Resources' scarcity narrows the possibility to improve services or to promote innovations, because of the duty to provide services without interruptions. At least a minimum amount of resource slack is a key driver for creativity and prompt reaction to problems.

Thus, the extractive model of interaction has the judiciary system itself as key value-setting reference. Relationships based on this model are mainly generated by Courts willing to take action and looking for support among stakeholder. The primary goal of this model is improving Court internal efficiency. The Court is in the position to choose the level of engagement and the areas of activity that need more support or organizational change.

A different model of interaction with organizational environment is possible: we define it here as *platform model* (Table 1).

In platform model the primary cause of interactions is public value and the will to contribute to the administration of socio-economic phenomena. The starting point is not a specific operational need, but rather the feeling of being an important subject in the inter-organizational network producing public value and solutions to social hardships.

Being the primary goal very different from the extractive model, the focus of relationship is not based on the judiciary system itself, but rather on public policy and social innovation sponsorship. This imply the possibility to engage with subjects that, until now, has hardly been used to get in relation with judiciary authorities. This also imply for Courts the will to engage with stakeholders in a relation between equals: Court does not select stakeholders but provides resources and knowledge that can be gathered by any subject interested.

The perimeter of collaborations is not determined. Major focus is placed on the right allocation of public resources and to the construction of a network able to analyse and implement effective policies in the region.

The focus on inter-organizational network is essential in the platform model. This new perspective is needed because of a contemporaneity shaped by high uncertainty and fast-changing scenarios, where nor the government nor the market alone seems to have a possibility to manage and solve societal problems [9].

Table 1. Extractive model and Platform model of interaction with stakeholders

Variables	Extractive model	Platform model
Reference values	Judiciary system values	Collaboration in the inter-organizational network managing public value
Primary goals	Operational critical issues Development of innovative service	Management of socio-economic phenomena
Actors	Selected within the judiciary system	All the actors interested in taking action
Scope	Restricted and defined by Court	Unrestricted Unfocused on Court's needs
Court's contribution	Limited willingness to change	Open data disclosure Willingness to collaborate
Actors' contribution	Operational resources Technical resources Adoption of innovative practices	Analyses of social challenges Public policies Project addressed to critical issues

4. Methods and Research Question

There are very few empirical contributions to open data field, focused mostly on evaluating policies and governments' actions to foster open data disclosure [1]; [10]. The empirical analysis proposed in this work is rather focused on organizational and managerial issues related to data disclosure in the judiciary system.

This study is not intended to prove the solidity and soundness of the platform interaction model that is an ideal type, a theoretical instrument supporting reasoning and research.

The research questions concern the understanding of the dynamics emerging when judiciary world meets open data government frame from an organizational systemic point of view. The case study offers the opportunity to dig into very important issues: which could be the best procedures to engage with open data? What are the main obstacles to the disclosure process? What are the drivers supporting Courts engaging the environment outside from their comfort zone?

The empirical case that will be presented in Section 5 has been investigated from a very peculiar perspective: as members of *Opendatagiustizia*[2] collective, we have been directly engaged in the project. From the methodological point of view we engaged in an action research [11].

It must be remarked that the case study is highly explorative: in Italy there hasn't been any attempt to introduce at the Court level the open data debate and open government actions. At the same time, there are initiatives that occurred for specific juris-

[2] Opendatagiustizia (http://opendatagiustizia.it/) is a research stream of C.O.Lab, the center for research and development of C.O.Gruppo Srl, in partnership with Associazione Ondata (http://ondata.it/).

dictions or at national level. For example, from 2016 the Constitutional Court provides archives with open data standards on its web site[3].

5. Empirical Case

Turin is a Metropolitan city in Northern Italy and its Court of Justice is one of the biggest in Italy. From several years Turin Court is considered as one of the most virtuous and innovative in Italy, developing organizational innovations in the relationship with citizens, in service design for lawyers and professionals, in monitoring procedural delays.

This innovation-friendly spirit is related with leadership, because of Presidents eager to take advantage of opportunities offered by PCT and digital tools. In addition, Court has been in the centre of a system in which key stakeholders, such as lawyer Association and bank foundations, contributed with economic support and resources to implement projects.

The first contact between Turin Court and Open Data was in the spring of 2017 in a public event connected with the International Open Data Day. In 2018 the appointment has been replicated with the will to start an explorative project in the Open Data field with the support and the endorsement of Court President and administrative chief.

Relevant figures of the administrative staff was contacted after a few weeks to carry on the work. Main occasions of interaction have been workshops held after the opening hours of offices.

The workshops were conceived to compose a typical *service design thinking* double diamond [12]. The adoption of design thinking framework was intended to treat open data as a service that Court could be able to provide. The process was structured in four moments: Discovery; Focus; Exploration and Operation.

People were separated into small teams to focus on a subject, taking account of specific competencies. During workshops, teams were asked to deal with canvases and reasoning schemes to deepen the knowledge of the subjects, the relation between procedures and data produced, the stakeholders involved and possible partners of the project.

The focus phase brought out the main themes sensible for further workshops activities. The most important themes were related to procedures involving forms of social hardship: the support of non-autonomous people; procedures related to broken families; evictions; financial difficulties of families and persons. More than this, a bunch of other themes emerged, more connected to Court administrative functions such as the list of registered professionals nominated by judges for impartial expertise.

The exploration phase had the role to transform hypothetic ideas in operative projects. Groups were asked to image how Court disposable data and activities could become meaningful for external organizations and citizens. At the same time, much attention has been placed on obstacles, on useful but not disposable data, on the issues related to privacy and on the resources needed to manage the disclosure and the update of open data.

Once the ideational work was done, groups dealt with the informative system database and test specific data mining strings to check the presence and the format of sensitive variables. Other work was needed to verify the consistency of data and to remove variables that could compromise in some way the anonymity of users and citizens.

[3] https://www.cortecostituzionale.it/jsp/consulta/rapporti_cittadini/open_data.do.

5.1. Operational Issues

Once the ideational side of the project was completed, the operational side raised several obstacles and pitfalls for the disclosure. Some of them were just snags for the provision of the most meaningful data, while others were quite conclusive.

The team working on family procedures faced the fact that a huge amount of information about families are stored in a template provided by the National statistics institute. The template is a handwritten paper form, therefore data aren't retrievable and useless for disclosure. The judiciary software stores quite exclusively information related to the procedure, avoiding social and economic data, because a double registration would require too much effort.

The team working on procedures supporting non-autonomous people decided to pay attention on georeferencing in order to map social hardship. The idea was to provide anonymous information through the ZIP code. The team found out that ZIP code was rarely filled in the database and other parts of the addresses were poorly filled. Georeferencing has been framed as the most important variable also by the team working on evictions but, also in this case, the ZIP code is never filled in the database.

Operational pitfalls can be considered unavoidable dealing with an explorative project. One of the main goals of the initiative can be considered an assessment of open data feasibility in the judiciary system and a full recognition of operative obstacles can be considered a first outcome of the project.

5.2. Systemic Considerations

Beyond operational issues, Open Data Giustizia Torino offered the opportunity to observe individual and organizational behaviour facing a potentially disruptive innovation. The case led us to investigate three levels of analysis that should be managed to implement a platform model of interaction with stakeholder and organizational environment: technological, bureaucratic and organizational, cultural and institutional.

5.2.1. Technology

In Italian judiciary system, the informatic system led to PCT reform and has been an instrument of organizational integration and efficiency within offices. Nonetheless, it has been designed to be as coherent as possible with offices traditional work. The informatic system has been mainly intended to reproduce paperwork on a digital support. Stretching system's potential exposes several elements that could restrain open data disclosure rather than easing it.

First, the system is not designed to manipulate and analyse data. It doesn't allow to have immediate account of statistics and metrics, and it doesn't easily enable business intelligence: information is not easy to pick up from the system and extractions are possible only through code strings.

Second, the insertion of data by offices is mainly about procedural issues. Data regarding individuals and families are often not taken in consideration. This makes sense for efficiency needs but prevents the system from being a social knowledge platform.

Third, some issues could be resolved by a direct connection with other public databases such as registry offices and tax agency, but judiciary system acts as a monad, rarely and insufficiently connecting to other public databases.

5.2.2. Bureaucracy and Organizational Structures

From a macro-organizational point of view, Courts' administration is focused on supporting the technological core [13] that can be identified in judges' decisions: the primary function of offices is channelling and shrinking uncertainty, together with formal controls imposed by law. Besides, Italian Courts administrative workers are in average quite old and difficult to stimulate because of laws regarding public workers. Very often the staff is underpowered.

These elements produce in Courts a strong focus on the contingent situation and on immediate operational needs. There is little room for strategic initiatives and long-term organizational goals. Organizational charts promote this situation, because, even in big and important Courts, there isn't an organizational role responsible for strategic long-term innovation programs and stakeholders' relationships, nor accountable of quality and quantity of data collected in the database with a systemic perspective. Italian judiciary organizations are very adherent to Weberian bureaucratic ideal type: this imply also little economic and organizational motives toward innovation and quality of work, because a public officer is hardly rewarded for producing effort toward organizational improvement and innovation.

In the Turin Court case, strategic organizational functions have been successfully held in last years by the President judge and the administrative chief, but a healthy organizational structure should be able to prevent too much pressure and commitment on individuals.

5.2.3. Organizational Culture and Institutional Logics

More than this, discussion can be carried on the ground of institutional logics and organizational culture.

Institutional logics have been defined as socially constructed values, beliefs, and rules, by which individuals constantly provide meaning to their reality, ordering material and social context [14]. Institutional logics operate at a macro-societal level but strongly influence organizational cultures: in the same organization there can be multiple institutional logics leaving space to diachronic processes of cultural change, to individual agency [15] triggered by institutional entrepreneurs [16] or competing logics [17].

Working in a Court means receiving a socialization based on a strong bureaucratic ideal type [18]: programmed and precise processes are the core of activities with little space to spontaneity and flexibility; formal rules define compulsory activities and legitimacy; hierarchy defines individual action space.

In relation to Courts' innovative processes, bureaucratic institutional logic implies that individuals have strong cognitive boundaries toward innovative practices. Everyday routinary activities are the only way to frame personal role, therefore a strong cognitive effort is required to reason on the Court systemic role on the territory in connection with stakeholders.

Given the bureaucratic organization, every office has a precise goal supporting organizational ends. Widening this frame to discuss how to reach general social goals, individuals could suffer the lack of the stable normal positioning.

Bureaucratic institutional logic influences also the way data are defined, treated and stored. In traditional way of working, administration of data is useful to report and certificate – e.g. to answer at an inspection. This strongly restrains the appreciation of knowledge produced by datification, mainly because an analytic approach to procedures is not required. There are strong impacts also on data quality, because data providing qualitative intensity and a better comprehension of social phenomena are often considered completely useless.

6. Conclusions

Our intent in this work is to stimulate the debate about the role of judiciary organizations in contemporary public administration's interaction with its environment in collaborative inter-organizational networks producing public value.

We frame this issue starting from some believes. First, public administrations – and particularly judiciary systems – has not yet understood how to make the best use of data produced by its own action. Second, the connection between judiciary system and stakeholders is crucial for quality and quantity of services provided, as demonstrated also by several elements of PCT reform implementation. Third, open data initiatives are a fundamental mean to energize and fill with meaning the relationship between judiciary organizations and their organizational environment.

Digital technologies offer several opportunities to engage with social environment, through the disclosing of data that can be meaningful to a large array of stakeholders. This is particularly true for Courts, given the wide magnitude of social and economic variables touched by their procedures.

In this sense, open data can be considered as a medium to establish new relationships and an innovative model of interaction with citizenship, that we described as platform model. Going further, analysis of data generated by judiciary organizations introduces other relevant issues: for example, artificial intelligence could be developed to provide instant insights on organizational effectiveness and social hardship emergencies.

In our opinion, the case study is particularly interesting because of several features. First, participatory methodologies are an instrument very effective and they should be spread more and more in the public sector: they offer the possibility to avoid enforcing top-down order that usually produce a formal and passive acceptance of new practices. Second, the disclosure process that has been described and analysed didn't derive from an imposition or a specific policy, but rather from the Court leadership willingness to find and explore innovative solutions. Third, the exploratory nature of the project predictably let emerge several obstacles and pitfalls.

It is necessary to highlight an important limitation of the case study: being the disclosure process *in fieri*, the project will produce outcomes mainly in the future. Nowadays, we don't know how many datasets will be disclosed, how disclosure will be managed at the organizational level, how the stakeholders will react to the disclosure, if the stakeholders will be ready to collaborate with the Court in order to reach more meaningful dataset, if there will be space for the creation of an effective collaborative inter-organizational network, for dialogue and shared reflection.

We hope there will be further occasions to test the platform model with empirical evidences. This is particularly true even because, as pointed out in some contributions [19], open data on its own has little intrinsic value, because the value is created by its use: platform model could be the infrastructure able to create value from data.

In conclusion, we would like to provide suggestions for further research and an invitation for the community of practice.

Research in open data disclosure processes should be focused in looking for case studies and best practices in which the disclosure produced a concrete improvement in stakeholders' engagement. It could be important to find cases in which the inter-organizational network produced systemic results, new policies, or social innovation initiatives.

On the other hand, from an organizational and administrative point of view, it could be primarily relevant to investigate the effects of open data disclosure on organizational behaviour and practices. It could be interesting to assess if workers would be able to develop a wider awareness of organizational systemic role and about the importance of data quality. More than this, it's very interesting the balance between strategic practices, innovative initiatives and bureaucratic requirements.

The invitation for the community of practice dealing with innovative solutions in the judiciary system is not to underestimate the potential relevance of open data in paving the way to an evolution of the interaction between the judiciary system and society at large.

Judiciary organizations could become an important catalyst for social innovation and data-driven innovative policies. To let this vision come true, it is necessary to explore new interactions with stakeholders and reframe in a strategic way several organizational practices. Opendatagiustizia group will gladly support all projects and experiences that will go in this direction.

References

[1] Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, 22.

[2] Kassen, M. (2013). A Promising Phenomenon of Open Data: A Case Study of the Chicago Open Data Project. *Government Information Quarterly 30*(4), 508-513.

[3] Bertot, J. C., Jaeger, P. T. & Grimes, J. M. (2010). Using ICTs To Create a Culture of Transparency: E-government and Social Media as Openness and Anti-corruption Tools for Societies. *Government Information Quarterly 27*(3), 264-271.

[4] Borzacchiello, M. T. & Craglia, M. (2013). Estimating Benefits of Spatial Data Infrastructures: A Case Study on e-Cadastres. *Computers, Environment and Urban Systems 41*, 276-288.

[5] Hood, C. (1991). A Public Management for All Seasons? *Public Administration 69*(1), 3-19.

[6] Osborne, S. P. (2006). The New Public Governance? *Public Management Review 8*(3), 377-387.

[7] Haveri, A. (2006). Complexity in Local Government Change: Limits to Rational Reforming. *Public Management Review 8*(10), 31-46.

[8] Stoker, G. (2006). Public Value Management: A New Narrative for Networked Governance? *American Review of Public Administration 36*(1), 41-57.

[9] Robinson, D. G., Yu, H., Zeller, W. P. & Felten, E. W. (2009). Government Data and the Invisible Hand. *Yale Journal of Law & Technology, 11*, 159.

[10] Zuiderwijk, A. & Jannsen, M. (2014). Open Data Policies, Their Implementation and Impact: A Framework for Comparison. *Government Information Quarterly, 31*(1), 17-29.

[11] Lewin, K. (1946). Action Research and Minority Problems. *Journal of Social Issues 2*(4), 34-46.

[12] Stickdorn, M. & Schneider, J. (2010). *This Is Service Design Thinking: Basics, Tools, Cases*. BIS Publishers.

[13] Thompson, J. D. (1967). *Organizations in Action: Social Science Bases of Administrative Theory*. Transaction Publishers.

[14] Thornton, P. H. & Ocasio, W. (2008). Institutional Logics. In Greenwood, R., Oliver, C., Sahlin, K. & Suddaby, R. *Handbook of Organizational Institutionalism.* Sage.

[15] Boltanski, L. & Thévenot, L. (2006). *On Justification: Economies of Worth*. Princeton University Press.

[16] Suddaby, R. & Greenwood, R. (2005). Rhetorical Strategies of Legitimacy. *Administrative Science Quarterly, 50*, 35-67.

[17] Scott, W. R., Ruef, M., Mendel, P. & Caronna, C. (2000). *Institutional Change and Health Care Organizations: From Professional Dominance to Managed Care*. University of Chicago Press.

[18] Weber, M., 1922 (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.

[19] Janssen, M., Charalabidis, Y. & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management, 29*(4), 258-268.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190028

Dealing with Privacy Issues in Data Integration: Scenarios for Official Statistics

Piero Demetrio FALORSI [a], Brunero LISEO [b] and Monica SCANNAPIECO [a]
[a] *ISTAT - Istituto Nazionale di Statistica (Italy)*
[b] *Sapienza Università di Roma (Italy)*

Abstract. The increase of data availability poses new challenges and suggests new interesting road to public and private data producers and providers. The European Commission acknowledged these opportunities that can significantly boost European competitiveness in the global market and in scientific research. One of the cornerstones of the process to build a common European data space is the possibility to access and share public and publicly funded data. This task has many important different goals: 1) citizens' secure access to and sharing of health data; 2) improving and innovating healthcare solutions based on mobile applications; 3) multiple uses of public sector information; 4) sharing scientific information, in order to facilitate the dissemination of results across countries; 5) Economics: Business to Business (B2B) data sharing, which considers the availability of "non personal machine-generated data". The new challenges suggest new problems to be faced both on a legislative and on a methodological ground. From a legal perspective, data exchange between public Institutions and private agents requires a detailed national legislative framework, still missing in many European countries. From a methodological perspective, the interaction between public and private data holders poses complex problems: 1) privacy preserving record linkage: how to guarantee that the linkage of personal data coming from different sources will not jeopardize the privacy of single citizens and/or companies; 2) the use of linked data as input to more sophisticated statistical analyses without unplanned information disclosure. Secure Multiparty Computation (SMC) techniques can play a role in this respect. In this Chapter we describe how the Italian National Institute of Statistics (Istat) is facing the new challenges and what are the most important steps to take in the next future.

Keywords. multiple sources, data protection, record linkage

1. Background: Official Statistics and National Statistical System

The role of official statistical national agencies has become more and more important in the last decades, especially in the most developed countries. The computational and methodological advances in the ability of managing, exchanging and combining different statistical information sources have dramatically increased the importance of such agencies which are going to play a central role in the process of stimulating data use and evidence based decision making by the governments.

The term ‘national statistical system’ (NSS) refers to a country’s producers of official statistics, generally a national statistical institute (NSI) and other institutions and administrations producing official statistics. National statistical system and, in particular, national statistics institutes (NSIs) have several duties.

In Italy, the Italian National Institute of Statistics (Istat) commitments are regulated by the so called National Statistical Program – also known as Sistan[1], which has a 3 years agenda – established with the Legislative Decree No. 322/1989, and it represents the legislative tool for planning statistical activities to be carried out by Sistan bodies, including Istat, in order to satisfy international standards and country’s information needs.

The National Statistical Program determines, for example, what are the statistical outputs which must be publicly available, to what level of detail the above statistical outputs need to be disseminated, in terms of variables, classification rule and aggregation level, in such a way to not cause disclosure risk and, at the same time, to guarantee a satisfactory information level. The National Statistical Program is divided into two volumes and an annex:

- The first volume is devoted to describe the way in which statistical information evolves and changes (framework information; information gaps by sector; aggregated costs of NSP).
- The second volume is much related to privacy issues and provide information and limitations about the activities processing personal data: it contains guidelines on personal, sensitive and/or judicial data processing.
- The annex is very technical: it establishes the methodologies which determine what level of disaggregation can be used for each single variable to be disseminated.

One of most relevant issues in the production of statistical information is the possibility of using administrative information *in lieu* of ad hoc statistical surveys. The reasons for doing that are multiple: administrative lists usually provide larger sample sizes, very reduced costs and minimal response burden.

On the other hand, they also imply a series of disadvantages, including measurement issues, and the fact that administrative data are generally not collected after an a-priori statistical design, their quality may be very low. In particular, they could not exactly answer the questions that a National Statistical Institute may want to ask.

A relatively new trend is to link both data sources, administrative records and survey data, to enhance the level of information and to open the way to more sophisticated data analyses.

It is not rare, however, that administrative data are collected by other public and/or private agencies and the operation of data exchange and linkage must be precisely regulated, especially at micro-level, where disclosure issues may arise.

Nowadays, in Italy, Istat can acquire administrative micro-data of public ownership for its mission. The entire data treasure consists, in 2018, of 478 administrative archives, owned by 94 different subjects.

The other subjects of Sistan may exchange micro-data from other subjects of the system which are strictly necessary for their specific mission.

Also, Sistan subjects are not allowed to release micro-data which are produced or acquired by others. There are a specific regulation and protocols for the supply of micro-

[1] www.sistan.it.

data for research. Even more complex are the relations between private agencies and Sistan members; in these cases it is necessary to build specific partnerships, in agreement with the data protection Authority.

2. Record Linkage Techniques and Connections with Privacy-preserving Issues

Today, the need for increasingly detailed and timely statistical information is shared by several international or supranational (European Union, European Central Bank, International Monetary Fund, etc.) and national (National Statistical Service, Ministries, Regions, etc.) bodies, and private users too. In this respect, increasing computational potential provides important opportunities. It is now possible to collect and maintain massive amounts of statistical data obtained from the integration of survey data and administrative information.

In this context, a significant problem is represented by the need to merge various data archives, possibly in view of the fulfillment of different goals. The awareness of the scientific community of this problem is testified by the numerous international symposia on "combining data from different sources". The main statistical approaches employed to address these problems can be classified as:

- Record linkage.
- Statistical matching.

The latter technique seeks to derive integrated statistical information by combining information from different datasets, in which only some variables are observed twice, and no overlapping of observed units is necessary. In this Chapter, we will only focus on the former approach. Record linkage refers to the use of specific algorithms that aim to identify pairs of records, corresponding to a single statistical unit, that are present in different databases. The same problem is addressed – albeit in a more general manner – in information technology literature, as the problem of integrating non-aggregated databases. In this context, relevant issues are (i) the construction of a general framework (ii) the detection and specification of semantic relationships between non- homogeneous data sources (iii) the characterization of data quality factors and (iv) the reconciliation of datasets from different sources, in order to construct a representation that is coherent with the relevant general framework and quality requirements. The following applications of record linkage methodologies deserve mention:

1. The construction and maintenance of a list of statistical units, to be used as a 'reference population' in sample or total surveys. In this context, it is important to identify units that feature in more than one database.
2. The merging of two or more databases to obtain a single archive, which is more informative at a non-aggregated level. This makes it possible to perform statistical analyses that would be otherwise impossible.
3. The use of several data sources for the improvement of the overall survey's 'coverage'. The methodological implications of these problems are not yet well-developed, but it is certain that the information provided by administrative data archives can be of great assistance in this regard.
4. Population size estimation problems via capture-recapture methods. A relevant example is the estimation of the under-coverage given by a complete census: this is usually performed via a linkage analysis between total survey data and an ad hoc post-enumeration survey [1].

5. The evaluation of the validity of a disclosure method, to protect access to administrative data from the risk of identification of single units by an intruder [2]; [3].

The use of record linkage techniques poses several interesting problems, in both methodological and computational terms. From the methodological point of view, the very definition of a statistical model (the description of how comparisons between records are performed) is still debated: see, for example, [4]; [5]; [6]; [7]; [8]. From a computational perspective, problems become formidable once the databases reach a large size (over 100 units). In these cases, comparisons are performed only between records that have the same values for certain 'blocking variables', which are assumed to have been recorded without errors. A broadly satisfactory solution of these problems appears, therefore, crucial.

In recent years, we have experienced a great proliferation of new Bayesian methodologies and, especially, an increasing number of statistical applications performed from a Bayesian perspective. The main reason for this trend lies in the development of Markov chain Monte Carlo (MCMC) methods which enable building and calibrating virtually any statistical model, regardless of complexity. This opportunity has made Bayesian methods much more appealing and visible in many areas of application, including official statistics. NSIs have several important and complex tasks; for their practical implementation, different kinds of – more or less – subjective operational decisions must be taken. For example, several important economic and social indexes are the result of procedures that at least implicitly involve the use of complex statistical models. Nevertheless, the result of a statistical analysis performed by NSIs 'must' be objective or, at least, should be perceived as such by users.

Bayesian concepts can be important for official statistics when (i) important prior (or extra- experimental) information on the variables of interest exists, and cannot be exploited adequately in a classical inference framework; and (ii) even when prior information is missing, a Bayesian analysis can be required, because a classical approach cannot provide answers unless strong assumptions, not easily tested, are introduced. In these situations, a Bayesian analysis enables at least a sensitivity analysis, to quantify the influence of the assumptions on the inferences made.

In general, from the point of view of statistical methodology, merging two (or more) data files can be important for two reasons:

- Per se, to obtain a larger and integrated reference dataset.
- To enable performance of a subsequent statistical analysis, based on the additional information obtained, that cannot be extracted from either of the two individual data files.

As already noticed, record linkage and disclosure techniques are intimately related because potential identification of units from record linkage techniques may disclose – even accidentally – sensible information at micro-level.

Consider a case of international crime investigation, where different databases from different countries and agencies are compared and linked to gain information. Similar examples arise in biomedical science where the use of integrated data may help in the detection of adverse drug reactions [9].

The disclosure and sharing of databases containing sensitive information is a very complex task and it must be regulated both from a legal and from a methodological perspective. Following [10], we define the privacy-preserving record linkage in the following way.

"*There are k different owners of databases D(1), ... , D(k). Each owner aims at determining which units in his/her databases match some units of other databases according to a decision criterion which compares strings belonging to different databases*".

Each owner does not wish to reveal his/her own actual records with any other party. They are willing to share with other parties, the actual values of some selected attributes of the record pairs which are classified as matches by the decision rule.

This can be done in several different ways. A first important distinction is between techniques which involve a third party or not. In a three party protocol, a third institution is involved in performing the linkage and it represents a filter between the two data owner. In a two-parties scenario, the two owners directly interact and sophisticated techniques are needed in order to avoid the disclosure of sensitive information during the linkage process.

3. Integration Scenarios: How to Preserve Privacy When combining Multiple Sources

In this Section we will highlight different scenarios where techniques like privacy-preserving record linkage could find application.

We envision four scenarios that can support the 'generic' information sharing need, namely: (i) private set intersection (PSI); (ii) private set intersection with enrichment (PSI-E); (iii) private set intersection with analytics (PSI-A); (iv) private data mining.

More specifically:

- Private Set Intersection (PSI): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact join to A and B without revealing any unnecessary information about their individual databases. That is, ideally, the only information learned by P1 about B is A∩B and vice versa (Figure 1).
- Private Set Intersection with Enrichment (PSI-E): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact or approximate join to A and B without revealing any unnecessary information about their individual databases. After that, they wish to enrich joined records with variables by both parties. At the end of the process P1 will learn additional P2 variables on A∩B and vice versa (Figure 2).
- Private Set Intersection with Analytics (PSI-A): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the intersection of A and B in a private way. At the end of the process, the only information learned by the parties (beyond the keys of the records belonging to the intersection) is the result of the analytics function (Figure 3).
- Private data mining (PDM): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the union of A and B without revealing any unnecessary information about their individual databases. At the end of the process, the only information learned by the parties is the result of the analytics function (Figure 4).

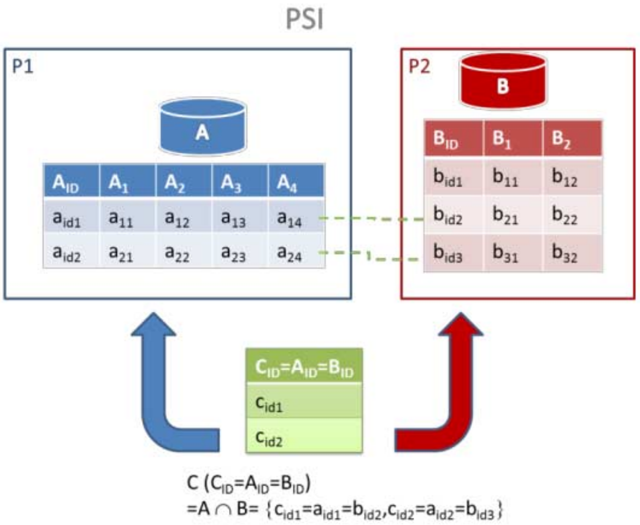


Figure 1. Example of PSI, P1 learnes that a_{id1} and a_{id2} are also owned by P2 and P2 learns that b_{id2} and b_{id3} are also owned by P1

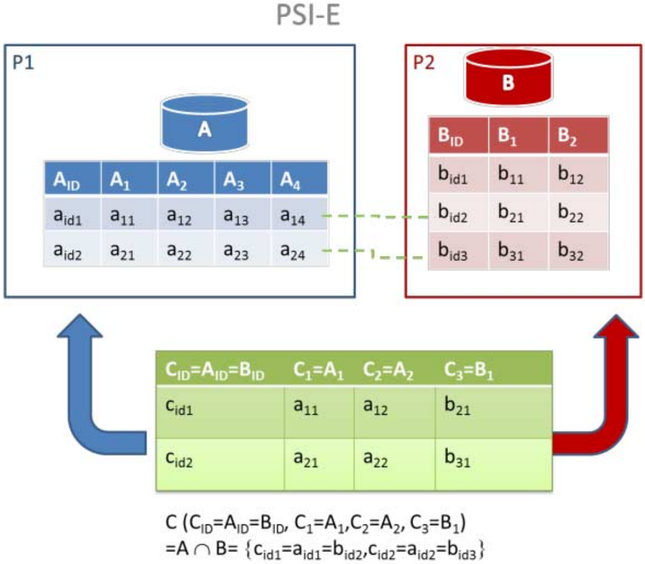
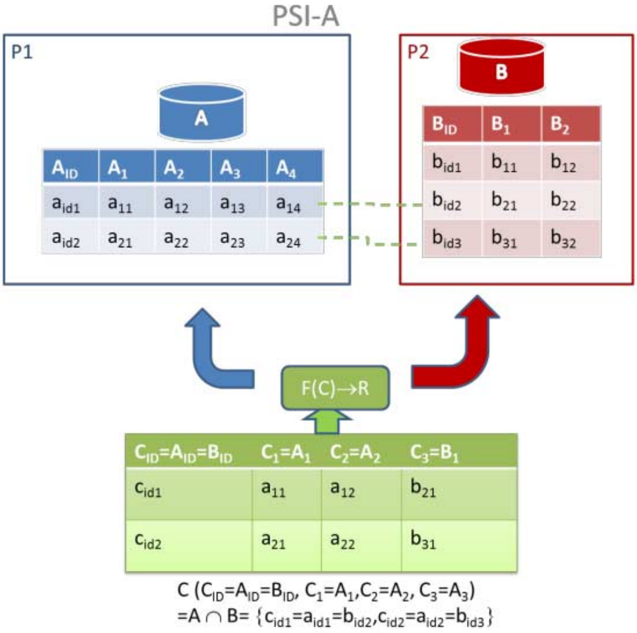

PSI-E
P1
A
A_ID A_1 A_2 A_3 A_4
a_id1 a_11 a_12 a_13 a_14
a_id2 a_21 a_22 a_23 a_24
P2
B
B_ID B_1 B_2
b_id1 b_11 b_12
b_id2 b_21 b_22
b_id3 b_31 b_32
C_ID=A_ID=B_ID C_1=A_1 C_2=A_2 C_3=B_1
c_id1 a_11 a_12 b_21
c_id2 a_21 a_22 b_31
C (C_ID=A_ID=B_ID, C_1=A_1,C_2=A_2, C_3=B_1)
=A ∩ B= {c_id1=a_id1=b_id2,c_id2=a_id2=b_id3}

Figure 2. Example of PSI-E, P1 learnes that a_{id1} and a_{id2} are also owned by P2 and also their values for attribute B_1. Similarly, P2 learns that b_{id2} and b_{id3} are also owned by P1 and also their values for attributes A_1 and A_2.



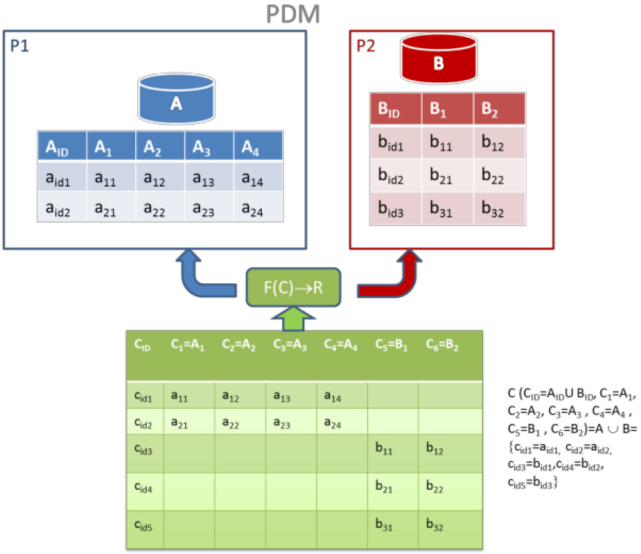
A_{ID}	A_1	A_2	A_3	A_4
a_{id1}	a_{11}	a_{12}	a_{13}	a_{14}
a_{id2}	a_{21}	a_{22}	a_{23}	a_{24}

B_{ID}	B_1	B_2
b_{id1}	b_{11}	b_{12}
b_{id2}	b_{21}	b_{22}
b_{id3}	b_{31}	b_{32}

$C_{ID}=A_{ID}=B_{ID}$	$C_1=A_1$	$C_2=A_2$	$C_3=B_1$
c_{id1}	a_{11}	a_{12}	b_{21}
c_{id2}	a_{21}	a_{22}	b_{31}

$C\ (C_{ID}=A_{ID}=B_{ID},\ C_1=A_1, C_2=A_2,\ C_3=A_3)$
$=A \cap B= \{c_{id1}=a_{id1}=b_{id2}, c_{id2}=a_{id2}=b_{id3}\}$

Figure 3. Example of PSI-A, P1 learnes the result R of the analytics function F applied to the intersection C and P2 gets the same



A_{ID}	A_1	A_2	A_3	A_4
a_{id1}	a_{11}	a_{12}	a_{13}	a_{14}
a_{id2}	a_{21}	a_{22}	a_{23}	a_{24}

B_{ID}	B_1	B_2
b_{id1}	b_{11}	b_{12}
b_{id2}	b_{21}	b_{22}
b_{id3}	b_{31}	b_{32}

C_{ID}	$C_1=A_1$	$C_2=A_2$	$C_3=A_3$	$C_4=A_4$	$C_5=B_1$	$C_6=B_2$
c_{id1}	a_{11}	a_{12}	a_{13}	a_{14}		
c_{id2}	a_{21}	a_{22}	a_{23}	a_{24}		
c_{id3}					b_{11}	b_{12}
c_{id4}					b_{21}	b_{22}
c_{id5}					b_{31}	b_{32}

$C\ (C_{ID}=A_{ID} \cup B_{ID}, C_1=A_1,$
$C_2=A_2, C_3=A_3, C_4=A_4,$
$C_5=B_1, C_6=B_2)=A \cup B=$
$\{c_{id1}=a_{id1}, c_{id2}=a_{id2},$
$c_{id3}=b_{id1}, c_{id4}=b_{id2},$
$c_{id5}=b_{id3}\}$

Figure 4. Example of PDM, P1 learnes the result R of the analytics function F applied to the union C and P2 gets the same

4. Conclusions

Even if Official Statistics already has a defined regulatory framework for privacy protection, the enforcement of privacy preserving measures at technical level in integration scenarios is necessary. In this Chapter, we have illustrated some specific scenarios for integrating data in a privacy-preserving way that could exploit techniques like privacy preserving record linkage. In order to implement solutions for these scenarios several aspects should be considered, namely: organizational, regulatory, methodological and technological. On the basis of our experience, one key factor is to have multidisciplinary teams working together on the specific objective. The investment from statistical organizations should be carefully considered and planned. However, there are many drivers that push for such investments, the main one being the fact that an organization can have access to data owned by another organization without directly accessing it, and without violating any privacy constraints. This fosters the flexibility of statistical organizations in answering statistical users' needs, while at the same time saving money and reducing response burden.

References

[1] Winkler, W. E. (1986). *Record Linkage of Business Lists*. Energy Information Administration, U.S. Dept. of Energy. Technical Report.

[2] Duncan, G. & Lambert, D. (1989). The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, *7*(2), 207-217.

[3] Winkler, W. E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, *Research in Official Statistics*, 1, 87-104.

[4] Fellegi, I. P. & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

[5] Copas, J. B. & Hilton, F. J. (1990). Record linkage: Statistical Models for Matching Computer Records. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 153*(3), 287-312.

[6] Belin, T. R. & Rubin, D. B. (1995). A Method for Calibrating False-match Rates in Record Linkage. *Journal of the American Statistical Association, 90*(430), 694-707.

[7] Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. (2001). On Bayesian Record Linkage. *Research in Official Statistics, 4*(1), 185-198.

[8] Tancredi, A. & Liseo, B. (2011). A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems. *The Annals of Applied Statistics*, *5*(2B), 1553-1585.

[9] Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1-16.

[10] Vatsalan, D., Christen, P., O'Keefe, C. M. & Verykios, V. S. (2014). An Evaluation Framework for Privacy-preserving Record Linkage. *Journal of Privacy and Confidentiality*, *6*(1), 35-75.

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190029

Opportunities and Challenges in the Legal Tech Services in the Italian and European Framework

Giuseppe VACIAGO
Università degli Studi dell'Insubria (Italy)

Abstract. After a succinct analysis of the legal tech sector, the aim of this Chapter is to highlight the main critical aspects as well as the evident opportunities which could arise from the introduction of these new tools throughout the legal market. The main critical aspects relate to the fairness and reliability of the algorithm, while there are opportunities to create an efficient, transparent and speedy process for receiving legal services which could translate into significant savings for public funds. The legal profession is at a critical juncture. Legal tech services, primarily involving the use of artificial intelligence, must not be overestimated, but neither should they be eschewed. The best policies must be introduced gradually because Italy and Europe are far behind the United States which has already gained considerable experience.

Keywords. legal tech, artificial intelligence, document automation, machine learning

1. Introduction

Over the last few years, the global legal tech market has grown exponentially, with a total of 3.81 billion US dollars invested between 2012 and 2018. The growth dynamics of investments in legal tech companies used to be low as investors were eyeing a fairly young business area and refrained from engaging in large transactions. In 2016, 224 million US dollars was invested in the industry; in 2017, 233 million US dollars was invested. 2018 saw explosive growth and investments became greater than ever, amounting to 1,663 billion US dollars [1]. However, there is a significant imbalance between the amount invested in the United States and the rest of the world: more than 3 billion US dollars in the United States, less than 200 million in Europe and less than 100 million in Canada. Interestingly, investment in Israel amounted to 12 million US dollars [2].

In Italy, legal tech is virtually non-existent with just a few companies included in the register of innovative start-ups at the Chamber of Commerce, whereas the total number at a global level is around 2,000. In any case, it is self-evident that the European market has not yet appreciated the potential for these services, unlike the United States where the common law system and the legal sector's enormous potential have fuelled a much more rapid increase.

Companies operating in the legal tech market offer different services to companies and law firms. With its CodeX[1] project, Stanford University in California has sought to divide companies into nine categories: 1. Marketplace; 2. Document Automation; 3. Practice Management; 4. Legal Research; 5. Legal Education; 6. Online Dispute Resolution; 7. E-Discovery; 8. Analytics; 9. Compliance. Currently the bulk of these companies provide Document Automation, Marketplace and Practice Management services but there is also significant growth in companies offering Compliance, Analytics and Legal Research services.

The various legal tech start-ups that fall into these nine categories must and will have to deal with the following 3 technologies that will inevitably change the legal sector.

First of all, the development of artificial intelligence has allowed some companies to promote software that can 'understand' a legal document. It is obvious that the level of understanding for now is limited to some very basic cognitive patterns which, however, for some sectors have already become essential. From this point of view, due diligence can be greatly helped by this software and some Italian law firms are already using software made in the United States to categorize and review a significant number of legal documents [3]. The same methodology can also be applied to the field of legal research with absolutely appreciable results [4]. The problem, as always in the field of artificial intelligence, is the quality and availability of the data to be analyzed. In Italy, for example, judgements are not always correctly digitalized in an interoperable format and the personal data present in these documents has not always been correctly anonymised. This makes it impossible to carry out an effective analysis in a legally correct manner that would allow the development of the predictive software already present in other jurisdictions.

Secondly, the use of blockchain in the legal sector will have and, in part, has already had a disruptive effect, even if the repercussions are not yet clear. However, the use of smart contract blockchain in complex markets such as international trade can ensure greater transparency, reduced costs and the guarantee of resolving any disputes in a much shorter time frame. The legal sector should work in the next years to employ and deploy smart contracts that could automate many functions. There is no doubt in fact that blockchain and smart contract are able to change our societal system and structures.

Thirdly, an extremely important and revolutionary aspect is legal design. Legal design has been defined as a human-centred approach to solving legal problems and legal innovation. This approach combines three different professional figures: the competence of the lawyer (legal), the mindset and methodologies of the designer (design) and, finally, the technological innovation of the computer engineer (tech) [5]. These three different figures have the possibility to create legal systems, services and processes, that are more usable, understandable and attractive. The role of legal design in the field of legal tech is fundamental, as the software has to standardize, as far as possible, particularly complex legal processes. To obtain this result, it's important to consider an approach which is geared towards clarity and transparency.

Given this succinct analysis of the related market and based on the practical experience gained by the author who recently set up a legal tech company[2], the aim of this

[1] CodeX, Stanford Center for Legal Informatics, https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics.

[2] The company is called LT42 S.r.l.

Chapter is to highlight the main critical aspects as well as the evident opportunities which could arise from the introduction of these new services throughout the legal sector.

2. The Main Critical Aspects of Legal Tech Services

2.1. Fairness and Transparency of the Algorithm

The first critical element stems from the risk that the algorithm may give false positives. One of the most dramatic examples is shown by the COMPAS system (Correctional Offender Management Profiling for Alternative Sanctions). Nationwide, many jurisdictions use statistical algorithms to assess the likelihood a defendant will fail to appear at trial or commit a future violent crime [6]. ProPublica analyzed 10,000 white and black defendants assigned scores in Broward County, Florida and they found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk [7]. More specifically: black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45% vs. 23%). At the same time, white defendants were often predicted to be less risky than they were. Propublica analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48% vs. 28%).

The false positive risk can also be found in other areas. St. George's Hospital in the UK developed an algorithm to sort medical school applicants. The algorithm was trained to mimic past admissions decisions made by humans [8]. But past decisions were biased against women and minorities. Whilst in the first example, the same rules determining the algorithm were by their very nature discriminatory, in the second case, the type of error committed is even more concerning as the algorithm codified discrimination.

The experiences described above concern borderline cases, but it is evident that whenever an algorithmic decision is applied to a legal process, problems can arise that impact the human rights and fundamental freedoms.

In this sense the European Regulation 679/16 (GDPR)[3] has clearly provided that a decision based solely on automated processing of personal data is admissible only if:

(a) it is necessary for entering into, or performance of, a contract between the data subject and a data controller;
(b) it is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
(c) it is based on the data subject's explicit consent.

This safeguard offered by the GDPR becomes an obstacle which is not easily manageable for those who promote the development of systems capable of making autonomous decisions.

The real challenge of the future is the transparency and fairness of the algorithm. However, when we think about the various tools offered on the market, the trend seems to be going in the opposite direction to that envisaged by the GDPR.

[3] Art. 22, EU Regulation 679/16 (Automated individual decision-making, including profiling).

2.2. Mida Paradox

The second critical aspect can be summarised as the 'Mida Paradox' according to which "Machines take the law literally... Humans don't" [9]. Within such a wide, complex panorama of information, it is quite possible that there are legal content losses in algorithmic translations of regulations. Any interpretation of the law presupposes not only knowledge of implicit information which is not contained within a document to be analyzed by artificial intelligence, but also the ability to perform assessments on the basis of a level of logical abstraction which it is difficult for a machine to attain. For example, any decision regarding a criminal defendant's mens rea on the basis of whether or not he wishes to commit a particular type of offence entails an extremely complex decision-making process which is very hard to translate into an algorithm. This paradox is increasingly fading. Artificial intelligence systems are undoubtedly developing mechanisms to avoid discriminatory situations and/or in violation of the GDPR, but there is still a long way to go.

In addition, it is important to consider that by inserting a layer of inscrutable, unintuitive, and statistically-derived code in between a human decision-maker and the consequences of that decision, AI disrupts our typical understanding of responsibility for choices gone wrong.

AI's introduces four concerns in terms of responsibility in the legal sector: 1) unforeseeability of specific errors that AI will make; 2) capacity limitations when humans interact with AI; 3) introducing AI-specific software vulnerabilities into decisions not previously mediated by software; and 4) distributional concerns based on AI's statistical nature and potential for bias. On the basis of these concerns described in the interesting paper written by Andrew Selbst, we could use the provocative term 'artificial negligence' to describe the risks of a decision taken by an automated process [10].

2.3. Digital Divide

The first two critical aspects illustrate just how important it is to invest in the legal tech sector, not only in financial terms but also in skills. The challenges faced by the Italian university system are the following: language barriers, lack of interdisciplinary approach, lack of legal practice and resistance to technological innovation in teaching methodology. Legal informatics courses should be significantly changed because they have never been more important as in this specific period to understand the future of law.

Starting from the assumption that in many Italian universities this course is not even compulsory, there are two alternative solutions: on the one hand, to add a specific course in legal tech, or, on the other hand, to significantly upgrade the current course in legal informatics. In addition to the traditional subjects of the legal informatics program (e.g. telematic process, digital signature) the following subjects should be taught: Big Data and machine learning, regulatory and ethical impacts of AI, coding for lawyers, smart-contracts and blockchain, computational law, e-discovery and legal design.

At a European level, the University of Helsinki offers a course in legal design and has created the Legal Tech Lab[4]; the University of Swansea offers an LLM in legal tech[5],

[4] http://www.legaltechlab.fi.

[5] http://www.swansea.ac.uk/postgraduate/taught/law/llmlegaltech/?utm_source=Artificial%20Lawyer&utm_medium=Legal%20Tech%20Education%20page.

while the University of Manchester offers a course entitled ‘Legal Tech and Access to Justice’[6]. At a national level, there are still no ad hoc university courses on legal tech, but there are important university initiatives that could lead to more structured courses[7].

If we don’t make these changes promptly, the risk would be twofold: on the one hand, to create legal tech services based on the illusion of the Mechanical Turk[8], whilst in reality legal work is still carried out by humans, on the other hand, to offer poor quality automated legal services which are not capable of meeting client needs.

3. The Main Opportunities Offered by Legal Tech Services

3.1. Access to Information

The computational capacity achieved in managing Big Data and machine learning systems makes it possible to provide the legal world with rafts of information which would have been inconceivable just a few years ago. Consider for example recidivism statistics or semantic analyses which can be run on judgements. Without wishing to enter the hazardous world of predictive justice [11], these types of instruments allow lawyers or judges to acquire a wealth of information which enables them respectively to defend their clients’ interests through in-depth legal research, and to hand down fair, correct decisions.

To give some concrete examples, Hague Institute for Innovation of Laws (HiiL) organized a competition entitled ‘Innovating Justice Challenge’ in 2018 which was attended by 33 African startups. ‘Gavel’, a Nigerian civic tech startup, aims to improve the pace of justice delivery through tech. Gavel does this by tracking criminal cases, and police brutality complaints. In addition, users are also connected to free legal aid lawyers through the platform[9]. Btrack Global is a Kenya startup that has developed tracking devices that enable motorcycle owners to easily track their motorcycles on their mobile phones in the event of theft[10].

These initiatives demonstrate that legal tech and access to information have a major social impact not only in United States or Europe, but also in countries with lower rates of technological skills. The solutions proposed by the African startups did not focus on the use of AI, but on the resolution of very concrete problems through web-based applications which are not particularly complex, but definitely very effective.

3.2. Greater Efficiency in Legal Services

It is undeniable that legal tech services can render a country’s legal system more efficient, providing greater transparency and rapidity in the decision-making process. These

[6] https://www.law.manchester.ac.uk/research/themes/law-money-technology/law-technology-initiative-2/.

[7] In addition to the initiative *Law Via the Internet* of the Institute of Theory and Techniques of Legal Information (ITTIG), the initiative *Technological Innovation and Law (TIL 2019)* and the excellent work done in the last years by Profs Palmirani, Sartor and Ziccardi deserve to be mentioned.

[8] The Mechanical Turk was a fake chess-playing machine constructed in the late 18th century. It was a mechanical illusion that allowed a human chess master hiding inside to operate the machine.

[9] http://www.gavel.ng.

[10] https://leap-2.com/en/projects/btrack-global.

services are however seen by traditional law firms as competitors. Nevertheless, legal tech services should not be viewed as competing with law firms, but instead should be conceived of as the future of the legal profession. It is beyond doubt that the digital revolution has changed many market sectors (consider for example the tourism or transport sectors). Despite this, the legal profession is still closely tied to a fiduciary relationship which cannot be affected by artificial intelligence, provided that there is willingness to accept that certain legal activities which are not particularly complex (for example a lease agreement for a building) must inevitably be automated. Once this type of approach has been accepted, the legal profession of the future will be essentially based on a lawyer's strategic ability to choose the best from amongst the various solutions put forward by artificial intelligence [12].

However, it is clear that the lawyer of the future will necessarily need to use these instruments with a level of competence that is not yet present at a national level. Access to information has always been a fundamental skill of the legal profession. The real challenge now will be to gain access to (profiled) information more quickly than one's competitors.

3.3. *Saving Public Funds*

One direct consequence of greater efficiency in a State's legal system is the opportunity to cut public expenditure. The introduction of the On-line Civil Trial system in Italy has enabled the state to achieve savings in the order of 178 million Euro over the last 3 years[11] and this result is only the start of the digitalisation process which will have beneficial effects on the entire judicial system.

One of the most frequent concerns of lawyers regarding the growth of legal tech is certainly to lower legal costs. However, this concern may turn into an opportunity: the lowering of legal fees could allow access to this type of service to a very different number of subjects (small enterprises, private citizens, non-profit associations, etc.) that currently cannot afford it.

Another concern is whether artificial intelligence will replace some legal services and consequently lead to job losses. A US study conducted in 2018 compared 20 well-respected corporate lawyers against an AI in an error-spotting test across a suite of non-disclosure agreements (NDAs). Responses were measured by time and accuracy. The human lawyers achieved an average accuracy of 85%, in an average time of 92 minutes. By comparison, the AI's success rate was measured at 92% in just 26 seconds[12]. The experiments carried out so far do not seem reassuring, but it's important to consider that drawing up an NDA or carrying out due diligence is an activity which does not obviate the need for subsequent legal advice which may be carried out with more time and resources.

[11] Report by the President Andrea Mascherin at the ceremony to inaugurate the National Forensic Council's 2018 judicial year, February 2018.

[12] Study conducted by LawGeek, one of the most famous legal tech companies in the United Kingdom, https://www.lawgeex.com/resources/AIvsLawyer/.

4. The Italian Scenario

After describing the opportunities and challenges faced by the legal-tech sector, an overview of the national scenario is appropriate, with a specific focus on the development of the most relevant technologies adopted in Italy.

As already mentioned above, Italy is lagging behind the rest of Europe, not to mention the United States. One of the main reasons is certainly the high number of lawyers practicing in Italy: in January 2018 there were 242,000 lawyers legally resident in Italy. On that date there were about 4 lawyers per 1,000 inhabitants. The income of a lawyer today is very different from 10 years ago, falling from an average annual income of 49,000 euros to 38,437 euros today[13].

The steady decrease in income prevents lawyers from investing in the technology sector. In addition, in Italy we prefer 'boutique firms', which in some cases, especially in the provinces, means that law firms are composed of a single lawyer or at most 2 or 3 lawyers who have not even formed a professional association.

The reduced spending capacity prevents lawyers from investing in human resources by hiring young lawyers who are often more familiar with technology. Furthermore, another is a widespread prejudice against the use of new technologies resulting from a real absence of a generational change as the increase in the retirement age and the reduction in income have deterred many lawyers from retiring.

4.1. Artificial Intelligence Tools

In this complex scenario, artificial intelligence tools are a 'luxury' that only a few law firms can afford, also because today there are few legal tech solutions based on artificial intelligence created specifically for the Italian market. Consequently, the tools that are sold globally can only be purchased by law firms that have an adequate budget and whose clients are large multinationals which can appreciate the type of output that these tools are able to achieve (especially in the field of due diligence).

Even though this type of scenario is certainly not reassuring, it does offer a great opportunity, many International legal tech companies are trying to establish synergies with Italian law firms to adapt their technological solutions to the Italian legal context. This type of activity can be very instructive especially for young lawyers who often find it difficult to find employment in law firms and who may instead have the opportunity to develop professionally in a different context.

Another aspect of fundamental importance for the growth of AI tools in Italy is undoubtedly the lack of availability of data in an interoperable format. From this point of view, it is necessary to consider that machine learning can only work in the presence of a large volume of data.

The adoption of On-line Civil Trials has allowed the creation of a remarkable number of judgments in digital format. Now it is necessary for this volume of data to be used by a machine learning system which can process it with respect for the privacy and fundamental rights of the individual. When this process is completed, there will be a Copernican revolution in the field of legal databases that are no longer able to meet the needs of lawyers. The real risk of this backwardness is that young lawyers find information more

[13] Statistics from 'Cassa Forense Nazionale' (Italian Previdential Institute of Lawyer).

easily on Google than on legal databases with a real risk of then providing incorrect or superficial legal advice.

4.2. Document Automation

If artificial intelligence is a sector where in Italy the legal tech sector has not yet achieved significant results, the document automation sector is starting to produce interesting software which is appreciated by the legal sector.

In recent years, the legal publishing sector has understood how important it is to invest in the IT sector and some compliance tools have been developed with particular reference to privacy after the coming into force of GDPR. The software houses already present in the professional sector (legal, accounting and audit) have certainly implemented investments, creating increasingly efficient tools for practice management, compliance and accounting.

One of the reasons for the success of document automation tools is the growth of software as a service (SaaS), which has allowed the development of tools that can be updated without any effort by the software house. The updating of a legal tech software poses two levels of complexity: the first, common to all software, is platform updating and bug fixing. The second is legal updating: one of the many challenges that the sector will have to face in the near future will certainly be the continuous changes to the legal framework. In Italy, an average of 500 laws are passed every year, for a total of 110,000 laws issued and currently in force. Every single change to an article of one of the 110,000 laws could dramatically change the functioning of legal-tech software. For this reason, the possibility of a remote software update makes it possible to provide a service which is always in line with the regulations in force. A stand-alone software, on the other hand, would require the user to install the new update with all the well-known compatibility problems if the operating system were to be changed.

The advantage of software as a service, however, is counterbalanced by the need for more attention to compliance with the GDPR, as Client data is stored in the cloud.

A further advantage of document automation software is the simplicity of implementation, because it does not involve a complex development activity as in the case of artificial intelligence. There are 4 steps to creating a software of this type: (i) preparation of the legal output (document, contract, report) that the user must be able to receive; (ii) creation of a check-list of questions in which the objective is to ask the user in a simple and clear way for the information necessary for the creation of the final document; (iii) preparation of a flow chart if it is necessary to insert a variable in the logical path to be followed by the user when providing the requested information; (iv) preparation of an interactive guide in order to facilitate the user in filling out the check-list.

Such software can be designed to facilitate the professional activity of a lawyer or can directly address the final Client (whether a company or a natural person) in order to allow the production of a document that can be used by the latter. The results of such a strategic choice are not trivial: in the first case, the software has the sole purpose of making the lawyer's work more efficient. In the second case, however, such tools could progressively take away the lawyer's professional activity. For the reasons already mentioned above in the AI field, it is very unlikely that document automation software can replace a lawyer, but it is legitimate to wonder what could happen when a lease or a complaint to the judicial authority can be drawn up through an automated check-list.

In this scenario, there are many questions to ask: what guarantee of legal reliability can a document generated by an automatic tool offer? Will the Judicial Authority be able to accept legal documents (think, for example, of a divorce) produced by an automatic system?

There are many questions, but one fact is certain: the transformation of a legal service into an automatic or semi-automatic process has begun and it is still too early to obtain a final balance of advantages and disadvantages.

5. Conclusions

The legal profession is self-evidently at a critical juncture. Legal tech services, primarily involving the use of artificial intelligence must not be overestimated, but neither should they be eschewed. The best policies must be introduced gradually because Italy and Europe are far behind the United States which has already gained considerable experience. Therefore, the essential first step in order to develop the legal profession of the future is - where appropriate - to transform the legal service into an organised process. Once this transformation has occurred, it will be possible to appreciate which aspects can be automated and which areas, out of necessity, must remain within the remit of professional lawyers.

A young US researcher recently wrote that "Robots Will Help Lawyers, Much Like Autopilot Helps Pilots. During the 1940s many pilots were afraid that they'd lose their jobs due to the rise of autopilot technology. That didn't happen though, as even 80 years later we still have pilots operating the airplanes even while autopilot helps them immensely" [13].

The future of our profession will not be very different: I imagine a scenario where a lawyer will have the task of monitoring the activity carried out by a computer, choosing from amongst the various options that the software will develop, the simplest, strategic and most convenient solution for the Client.

References

[1] Pivovarov, V. (2019). 713% Growth: Legal Tech Set an Investment Record in 2018. *Forbes*, https://www.forbes.com/sites/valentinpivovarov/2019/01/15/legaltechinvestment2018/#6be33d27c2ba.

[2] Gupta, D. (2017). *Feed Report Legal Tech*. Tracxn Technologies Research.

[3] Alarie, B., Niblett, A. & Yoon, A. H. (2018). How Artificial Intelligence Will Affect the Practice of Law. *University of Toronto Law Journal, 68*, 106.

[4] Baker, J. J. (2018). A Legal Research Odyssey: Artificial Intelligence As Disruptor. *Law Library Journal, 5*, 110.

[5] Hagan, M. (2017). *Law + Design Workbook*, http://www.legaltechdesign.com/2017/10/law-design-workbook/#16.

[6] Jung, J., Concannon, C., Shroff, R., Goel, S. & Goldstein, D. G. (2017). Simple Rules for Complex Decisions. *Harward Business Review*.

[7] Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). *Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*. ProPublica.

[8] Gholipour, B. (2018). We Need to Open the AI Black Box Before It's Too Late. *Futurism*.

[9] Goel, S. (2018). Fairness, Accountability, and Transparency of Algorithms. *CodeX Future Law Conference*.

[10] Selbst, A. D. (2019). Negligence and AI'S Human Users. *Boston University Law Review, 100.*

[11] Ferguson, A. G. (2017). Policing Predictive Policing. *Washington University Law Review, 94*(5), 1109.

[12] Susskind, R. (2017). *Tomorrow's Lawyers: An Introduction To Your Future Paperback*. Oxford University Press, 120.

[13] Wolf, R. (2018). Why AI Could Never Substitute a Lawyer. *Law Technology Today*, https://www.lawtechnologytoday.org/2018/01/why-ai-could-never-substitute-a-lawyer/.

This page intentionally left blank

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

Subject Index

Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

Author Index

This page intentionally left blank

This page intentionally left blank