# From Legal Documents to Legal Document Management Systems; The Case of LegiCrowd

Alexandros Nousias
Future Now Business Consultants & Training / MyData Greece)
Athens, Greece
alexandros.nousias@gmail.com

Alain Couillault
Association des Professionnels des Industries de la langue (APIL)
Montreuil, France
alain.couillault@apoliade.com

Sofia Almpani
National Technical University of Athens
Zografou, Greece
salmpani@mail.ntua.gr

Theodoros Mitsikas
National Technical University of Athens
Zografou, Greece
mitsikas@central.ntua.gr

Petros Stefaneas
National Technical University of Athens
Zografou, Greece
petros@math.ntua.gr

## ABSTRACT

In this position paper, we argue that users' online consents to terms of services and privacy notices is naturally impaired by the unbalanced powers between online service providers and their users. We argue that a full fledged legal document management system relying on semantic representation is key to resolving this conflict and facilitating transparency of Online Legal Documents, and we give a quick overview of the LegiCrowd project, a crowdsourced approach to legal documents annotation, which paves the way towards such solution.

## 1 INTRODUCTION

As AI technology and automation permeate society horizontally, the law and the subsequent enforcement mechanism prove incapable of keeping pace. Concepts originating from the past like consent tend to maintain their static properties in an increasingly complex and dynamic space, thus resulting in a state of obsolescence. The law and its design and implementation properties are in need of radical update. The present paper argues that such update requires a transition from plain legal text to a full-fledged Legal Document Management Systems.

World Wide Web today is the outcome of a three stage evolution. Web 1.0 refers to the so called static Web of documents in a unidirectional broadcasting format. Web 2.0 introduced the web of people, by allowing the sharing of user generated content and further social networking. Web 3.0 or the Web of data, is currently evolving under the idea of defining and linking structured data [1] in order to produce formal semantic representations thus introducing massive automation via algorithmically informed decisions. The Web 3.0 comes however with one major loophole; the lack of legal knowledge modelling and representation, which emerges systemic inadequacies in the digital design, as the always hungry-for-data service supply side conducts a "permissionless invasion"[8]. However, in a complex dynamic system like the Web of data, algorithms require huge amounts of high quality and relevant data. We start from the basics revisiting the concept, role, and specs of terms of services and privacy policies as agents of information provision towards systemic, human centric, and human friendly automation. Terms of service and privacy policies are deemed raw data for automated meaning extractions via relevant information retrieval, question answering, dialogue systems, and other Natural Language Processing applications.

The rest of the paper is organised as follows: In Section 2 we provide a brief description of the information technology advancements to date and key characteristics thereof. Section 3 discusses inconsistencies and loopholes of the modern legal design. This Section also expands on that ground we argue that legal representation and modelling could be the solution for a radical update of the modern legal properties and enforcement mechanism, if put in the appropriate ethical context. Section 4 introduces the LegiCrowd platform, a crowdsourced legal document annotation system. Finally, Section 5 concludes the paper and provides some thoughts for future work.

## 2 FROM LEGAL TEXT TO LEGAL INFORMATICS

Ubiquitous automation does not support the static format of the online legal documents and the linked consent models. Terms of services and privacy policies in their present form constitute an iconic proof of inadequacy of the digital design. Complicated legal and technical documents that no one reads, no one understands, and no one cares about, govern the emerging data lifecycles for the benefit of data driven business operations by extending their unhealthy operational patterns. A piece of information of such magnitude turns into an irrelevant node in the data value chain, hindering the unfolding and systemic assertion of the evolving human centric patterns. On top of that, modern businesses increasingly use consent as a *de facto* standard for demonstrating privacy commitments and wider legal compliance claiming consent provisions as proxies of informed choice. This evolution has given rise to a situation where many technology giants, on the pretext of providing improved services, have begun to track every action of every user with little or no transparency [6]. The result has been that clicking the 'Agree' button for consent was dubbed "the Biggest lie on the Internet"[7]

*WAIEL2020, September 3, 2020, Athens, Greece*
Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2844/ethics6.pdf

and incidents of data misuse such as unsolicited call, spam and deliberate manipulation have resulted in a massive trust deficit. And all that formalised by the court's validation of the 'I Agree' button maximising the power asymmetries and the trust deficit.

## 3 THE ETHICS OF LEGAL REPRESENTATION AND MODELLING

Such a formatting reality and the imposed data dispossession from the technology and digital service supply side, brings into the surface the need for dynamic, data-driven, and data-relevant legal and ethical enforcement. In the environment of Web 3.0, such an enforcement requires a data driven solution shaped with mathematical reasoning. It requires the transition to a ubiquitous legal representation and modelling apparatus; an extended Legal Document Management System comprised by structured legal data, methods, and tools for sufficient syntactic and semantic representation, capable of generating documented, machine readable legal knowledge, using very different logic, norms, and languages.

The ethical starting point lies on the axiom expressed by [2] that "*The common misconception is that language has to do with words and what they mean. It doesn't. It has to do with people and what they mean*". It is not about simple language data linking and annotation, rather about **providing accurate meaning in the appropriate context**. The aim is a virtuous cycle of legal data structuring, modelling, representation and context in order to: (i) Provide end users spot on clear and ascertained information on data processes and circulation; (ii) Provide the supply side proof of concept for technical and legal compliance throughout the data lifecycle, thus mitigating compliance inconsistencies and pertaining risks; (iii) Turn to a standard design building block; (iv) Enhance platform transparency and user confidence and trust; (v) Embed into the increasing B2B, B2C, C2C as well as Device to Device (D2D) data flows ethical requirements, like human agency and oversight, technical robustness and safety, privacy and data governance, (OLDs) fairness, accountability, etc.

## 4 THE LEGICROWD APPROACH

The LegiCrowd project could be an answer for such a need for transparency, as it aims at creating a platform to render Online Legal Documents (OLDs), namely Privacy Notices and Terms of services, in a quick and easy to read format, such as icons, dataviz or simplified language through a crowdsourced approach. This requires first to design a semantically sound annotation tag set, as an ontology of descriptors. This is the goal of the current LegiCrowd Onto project, which relies on a number of competencies particularly related to natural knowledge modelling, law and corresponding visualisations thereof gathered in an international consortium. Such a platform aims at truly putting end users in the driver's seat as it a) provides an ethical building block in the overall design, b) empowers end users to extract accurate legal information in context, to assess the levels of legal compliance and the ethics standards in place and c) provide or reject a consent on a truly informed basis.

## 5 CONCLUSION

No doubt, the practice and assertion of law in the Web 3.0 era is a combination of numerous language data inputs and outputs from multiple workflows. In the said legal workflows, the extraction, formulation, and exploitation of related metadata and provenance constitute a basic processing component towards Machine Learning models or Natural Language Processing applications, capable for more efficient legal enforcement. With high awareness of its potential societal impact, any decisions about legal data, methods, and tools tend to tie up with their impact on people and the society in a practical way thus bringing ethics in the automation foreground.

## ACKNOWLEDGMENTS

The LegiCrowd Onto consortium is lead by the French Non Profit Organisation Association des Professionnels des Industries de la Langue (APIL), and includes the National Technical University of Athens (NTUA) and the Research, Consultant & Training firm 'Future Now', backed by MyData Greece, the Greek node of MyData Global. It has received funding from the European Union's Horizon 2020 research and innovation programme under the NGI TRUST grant agreement no 825618. This project has been made possible thanks to Short Term Scientific Missions conducted within the framework of the enet collect Cost Action ([3], [4], [5]).

## REFERENCES

[1] Nupur Choudhury. 2014. World Wide Web and Its Journey from Web 1.0 to Web 4.0.
[2] Herbert H. Clarck and Michael F. Schober. 1992. *Questions about question - Enquiries into the cognitive bases of surveys.* Russell Sage Foundation - New York, New York, NY, USA, Chapter Asking questions and influencing answers, 15–48.
[3] Alain Couillault. 18/5/2018. *SHORT TERM SCIENTIFIC MISSION (STSM)SCIENTIFIC REPORT.* Technical Report. Apoliade. http://www.enetcollect.net/ilias/goto.php?target=file_530_download
[4] Alain Couillault. 3/3/2019. *SHORT TERM SCIENTIFIC MISSION (STSM)SCIENTIFIC REPORT.* Technical Report. Apoliade. http://www.enetcollect.net/ilias/goto.php?target=file_908_download
[5] Alain Couillault. 8/3/2020. *SHORT TERM SCIENTIFIC MISSION (STSM)SCIENTIFIC REPORT.* Technical Report. Apoliade. http://www.enetcollect.net/ilias/goto.php?target=file_1053_download
[6] Joss Langford, Antti Jogi Poikola, Wil Janssen, Viivi Lähteenoja, and Marlies Rikken. 2019. *Understanding Mydata Operators.* Technical Report. MyData.org.
[7] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.
[8] Tom Wheeler. 2018. Time to Fix It: Developing Rules for Internet Capitalism. *Fellows Research Paper Series. Shorenstein Center on Media, Politics and Public Policy* (2018).