From RELAW Research to Practice: Reflections on an Ongoing Technology Transfer Project

Nicolas Sannier, Mehrdad Sabetzadeh, Lionel C. Briand
SnT Centre for Security, Reliability and Trust
University of Luxembourg, Luxembourg
nicolas.sannier@uni.lu, mehrdad.sabetzadeh@uni.lu, lionel.briand@uni.lu

***Abstract*—Over the past years, we have been studying the topic of automated metadata extraction from legal texts. While our research has been motivated primarily by RE problems, we have observed that the interdisciplinarity of the research on legal metadata, and indeed on several other topics considered by the RELAW community, has the potential to trigger innovation beyond the traditional RE. In particular, legal metadata is a key enabler for the rapidly-expanding field of Legal Technology (LegalTech). In this short paper, we describe the preliminary steps we have taken toward transitioning a prototype tool for legal metadata extraction (developed in our previous work [1]) into a platform that is palatable to the LegalTech market. We hope that our findings would provide useful insights about the value chain for legal metadata and further offer a concrete example of a technology transfer attempt that is rooted in RELAW research.**

***Index Terms*—Legal Metadata, Technology Transfer.**

I. Introduction

Legal Technology (LegalTech) is concerned with applying information technology for assisting legal professionals with their day-to-day tasks, including, among others, content management, legal search and discovery, regulatory monitoring, reporting, and compliance checking [2]. In recent years, LegalTech has also taken under its umbrella software-based solutions that are aimed at improving the traditional legal practice by either reducing the need for legal human resources or using these resources in a more efficient manner [3].

An important prerequisite for many aspects of LegalTech is to have the structural and semantic properties of legal texts expressed in an explicit and machine-analyzable form. These properties constitute the *metadata* that needs to be recorded alongside the natural-language content of legal texts, both to facilitate the interpretation of legal texts by humans and to enable advanced automated analysis.

Given the sheer scale of legal corpora written over decades and centuries, a fully manual creation of legal metadata is laborious and requires a tremendous amount of resources. In the RE community, there has been considerable research on automated extraction of metadata from legal texts. This metadata covers structural information, for example, the hierarchical organization of legal texts and the cross references (citations) in them [4], [5], as well as semantic information, for example, rights, permissions and obligations [6], [7].

The above-cited work has naturally focused on the interplay between legal metadata and legal requirements. This focus has helped keep the body of research cohesive and concentrated on a set of common RE goals. At the same time, we believe that this focus has made it more difficult to build a "critical mass" of industrial interest around the research outcomes. Building such a critical mass is important if one wants to break the technology transfer barrier and justify the high cost of maturing research prototypes into tools that can be applied in production environments.

In this short paper, we describe the preliminary steps we have taken toward increasing the industry readiness of a legal metadata extractor that we developed in our earlier work [1]. While our research activities were originally motivated almost exclusively by RE topics, over time and mainly due to the critical mass issue noted above, we came to the realization that we needed to expand the scope of our activities in such a way that we could respond to the broader needs of the LegalTech industry. We believe that this expansion of scope is beneficial to RELAW. In particular, the objectives pursued by RELAW are part of a larger set of objectives from a much larger community that is interested in the same or similar technologies. Being more aware of what this larger community does and needs is important, as this will enable RELAW to (1) tackle an even richer set of research problems, and (2) target larger groups of stakeholders for technology transfer and commercialization.

What we describe here is an account of work in progress. Our findings to date center around the following two research questions (RQs):

RQ1. (Aside from legal requirements analysts,) who are the main beneficiaries of legal metadata in the LegalTech market?

RQ2. What key qualities and requirements should one consider when building a minimum viable product for automated metadata extraction?

In the remainder of the paper, we discuss the above RQs (Sections II and III), and outline the next steps of our work (Section IV).

II. Market Needs

Below, we describe our findings about the market landscape for legal metadata, covering the observed needs from different stakeholders' perspectives. We elaborate market needs along four dimensions: (1) governmental entities, (2) professional

legal publishers, (3) companies specializing in textual content classification and analytics, and (4) companies that provide automation services for legal compliance, for example, to corporate law.

Governments. An important thrust of e-Government is increasing citizens' access to legal texts such as laws and regulations through online portals. While formats such as HTML or PDF have traditionally been the norm for these portals, there has been a rapid shift in recent years toward markup representations that provide legal metadata alongside the texts. Examples of portals that offer legal metadata include LegiFrance in France (http://legifrance.gouv.fr), BelgiumLex in Belgium (http://www.belgielex.be/), Overheid in the Netherlands (http://overheid.nl/), Legilux in Luxembourg (http://www.legilux.public.lu) and the Eur-Lex portal for the European Union (http://eur-lex.europa.eu).

An important challenge that governments face with regard to the provision of legal metadata is the large volume of legacy texts that contain no systematic metadata. Manually enhancing these texts with metadata is extremely time-consuming. Despite this, our interactions with several governmental entities indicate that this task is still done largely manually. Consequently, important compromises have had to be made in all the legal portals mentioned above, in order to reduce the costs associated with legal metadata. For example, the metadata about the structure of legal texts would typically go down only to the level of articles, without distinguishing the articles' subdivisions such as paragraphs, numbers and sentences.

Another major limitation in the existing portals that we have observed is that the cross references (citations) are not systematically resolved. Stated otherwise, the dependencies between different legal provisions are not currently captured in the metadata. Consequently, no automated support can be developed for analyzing the impact of the amendments made to legal texts. Currently, legal drafters incur substantial effort over manually analyzing the impact of amendments. This task is not only laborious but also error-prone. Inconsistencies in citations, for example, references to provisions that have been repealed, are not uncommon. Such inconsistencies and the legal ambiguities that may ensue can cause major practical issues, including loopholes and potentially economic loss.

Recent initiatives on visual authoring tools for legal texts, for example, LIME [8] and LEOS [9], are likely to reduce the manual effort associated with adding metadata to new or existing texts. Nevertheless, due to the sheer scale of the legal corpora for which systematic metadata does not exist, there is still a great need for automated retrieval of legal metadata. Such automation can reduce operational costs and, at the same time, provide richer and more accurate legal metadata.

Legal Publishing. There are a variety of commercial services in the legal domain revolving around the provision of legal insights, interpretation of laws and regulations, and tutorials and training for lawyers. These services are often specialized according to topic, for example, for tax and accounting, international rights, governance, human resources, criminal procedure, health, and so on. In this area, there exist numerous commercial legal publishers (increasingly also referred to as legal service providers due to their expanded operations, as we discuss below) who have a dedicated focus on developing annotated editions of legal texts, compiling jurisprudence, and offering legal training. Some major international players in this market segment are: LexisNexis (http://www.lexisnexis.com/) (US), Thompson Reuters (http://legalsolutions.thomsonreuters.com/) (US), RELX (previously Reed Elsevier) (http://www.relx.com) (UK) and Wolters Kluwer (http://wolterskluwer.com) (NL). There are in addition many national publishers such as Dalloz (http://www.editions-dalloz.fr) which acts as the de-facto national reference in France.

Until recently, the core business of legal publishers was centered around the publication and dissemination of physical books, which were updated on a regular basis. Legal publishers have now further developed a substantial footprint in online services, based on thematic subscriptions. These subscriptions usually encompass computer-supported legal search and can be complemented with other professional legal services. Legal search heavily relies on metadata in order to provide smart facilities such as e-discovery [10]. Both the physical books and the online subscriptions are very expensive, in part due to the high degree of manual effort spent by the publishers over annotating and linking legal texts. One can therefore easily surmise that legal publishers will gain a lot of competitive advantage through automated metadata extraction from legal texts.

Content Classification and Analytics. Besides the legal publishers, who dominate the legal information market, there are a number of smaller businesses whose activities are oriented around supporting companies that have to deal with a mix of internal data and legal information. In this area, automated classification, document management, smart search and analytics over such content as emails, contracts, and meeting reports can significantly boost not only productivity but also the level of legal compliance.

Classification and analytics services are based on a myriad of integrated applications and technologies, including optical character recognition (OCR) for managing legacy content, data storage, indexing and search, data mining, machine learning, templates, and access control policies. All of these applications and technologies rely to varying degrees on metadata. These applications are indeed often lightweight, web-/cloud-based successors to much heavier-weight workflow management and enterprise resource planning solutions. In this area, robust automated metadata extraction can offer substantial value for classifying legal content or content that is subject to legal constraints.

Legal Compliance Automation. This market segment is closely-related to the previous one, that is, Content Classification and Analytics. Nevertheless, compared to the previous market segment, legal compliance automation has a more specialized focus on the legal implications of the operations of companies that are regulated. In particular, the regulatory

requirements and rules that need to be automatically enforced here are not targeted at business efficiency and productivity, but rather at providing compliance evidence and running automated compliance checks. It is important to note that the broad field of legal compliance is already heavily dominated by major auditing firms, notably the "Big Four" (Deloitte, PwC, KPMG and EY). Nevertheless, alongside the auditing firms, a smaller but thriving compliance service market is emerging. This smaller market is targeted at catering to the needs of the following: (1) small companies that are subject to legal constraints, but which do not need the comprehensive (and often expensive) services of major auditing firms, and (2) companies with highly-nuanced compliance needs which cannot be adequately met by the general auditing and advisory services provided by major firms. Success in the legal compliance automation market rests on the ability to provide high-quality (compliance) services at lowered costs. A critical facet of automation in this context is to be able to (semi-)automatically derive executable rules for assessing and estimating compliance. Legal metadata plays an important role here as an intermediate step for building executable compliance rules.

Market Summary. The market segments outlined above have different needs with regard to legal metadata. These needs include, among others, supporting legal search and e-discovery, synthesizing legal interpretation and insights, developing automated compliance checking rules, and handling evolving laws and regulations. Most LegalTech applications use legal metadata under the hood. At the same time, little tooling exists at the moment that addresses legal metadata extraction in an explicit manner. From our discussions with various stakeholders, there appears to be a healthy demand for tool support around automated legal metadata extraction.

III. Toward an Industrial Legal Metadata Extraction Framework

A. Desired Qualities of the Framework

From our preliminary investigation, we have elicited a list of qualities that an industry-strength metadata extraction framework should possess. These qualities are:

Q1. Ability to automatically extract accurate metadata *(Effectiveness)*.

Q2. Ability to be tailored to the specific nuances of legal texts in different countries and different legal jurisdictions *(Customizability)*.

Q3. Ability to handle large legal corpora *(Scalability)*.

Q4. Quality of being easily learnable and applicable by the target users (*Usability*), particularly taking into consideration the customization facilities that the framework needs to offer in order to meet Q2 above.

Q5. Quality of being integrable *(Integrability)*, particularly into existing solutions that potential customers may have already invested into.

To meet the above qualities, we believe that a legal metadata extraction framework needs to employ a combination of Natural Language Processing (NLP) and conceptual modeling. A wide range of NLP techniques are pertinent, including, lexical analysis, phrase detection, dependency extraction, and natural-language similarity metrics. In our vision, conceptual modeling will enable the precise specification of concepts that are required to guide automation and analysis in a given context. Models can be used, for example, to make explicit the organization of legal texts as well as possible variations in the organization. Furthermore, operational code for metadata extraction can be derived via a variety of automated model-to-code transformation technologies. More specifically, models, via automated transformation, will generate the code that tunes and invokes the appropriate NLP modules for text processing. In this way, a user will be able to conveniently customize the framework for their needs with minimal exposure to software code and sophisticated cascades of NLP modules.

Depending on the application context, the extracted metadata may be used for different purposes, for example, the generation of legal portals or rule-based compliance analysis, as noted earlier.

We next present the core requirements we foresee for an industrial framework aimed at legal metadata extraction.

B. General Requirements for the Framework

We define seven general requirements that we believe are essential for an industrially-successful legal metadata extraction framework.

R1: Fine-grained segmentation of legal texts. The framework shall be able to automatically and precisely segment legal texts into their constituent parts, going from high-level divisions, such as books, chapters, all the way down to articles and article subdivisions, such as paragraphs, alineas, and lists. This requirement is a prerequisite for further automated text analysis.

R2: Advanced cross reference handling. Substantial work has already been done on the analysis of cross references [1], [4], [11]. The framework shall be able to leverage the extensive academic research already performed and incorporate means for automatically detecting and resolving cross references. This requirement will not only enable navigation within and across legal texts, but also serve as a basis for conducting change impact analysis.

R3: Semantic metadata extraction. The framework shall be able to support the extraction of semantic legal metadata. Semantic legal metadata has received a lot of attention in the RELAW community. Some notable semantic legal metadata items include modalities (such as rights, obligations and permissions) [7], [12], [6], [13], [14], conditions, consequence, and intent [15], [4], [11], [16].

With regard to the development of an industrial metadata extraction framework, the importance of semantic metadata originates primarily from the increasing interest in rule-based laws and regulations. Our interactions with governmental entities in Europe indicate that rule-based laws and regulations are already being considered as a mechanism for increasing

TABLE I
CONTRIBUTIONS OF REQUIREMENTS TO THE DESIRED QUALITIES

	Effectiveness (Q1)	*Customizability (Q2)*	*Scalability (Q3)*	*Usability (Q4)*	*Integrability (Q5)*
Fine-grained segmentation of legal texts (R1)	✓				
Advanced cross reference handling (R2)	✓				
Semantic metadata extraction (R3)	✓				
Metadata editing and visualization (R4)	✓			✓	
Customization facilities (R5)		✓	✓	✓	✓
Versatile NLP engine (R6)		✓	✓		✓
Standardized markup format for legal metadata (R7)		✓		✓	✓

transparency, supporting complex legal search, and facilitating the transition from law to compliant IT systems, for example, tax systems in the domain of public administration.

R4: Metadata editing and visualization. The framework shall be able to provide editing and visualization facilities for legal metadata. A user interface (UI) represents an important usability / acceptability criterion. The metadata extraction prototype we developed in our previous work [1] employs the annotation editing and visualization UI that is provided by the underlying NLP engine (see F6, below). While practical for software development purposes, this UI is insufficient and also too complex to be used by end-users, noting that the end-users include legal experts who may have little familiarity with software development. Consequently, UI issues need to be investigated more deeply.

R5: Customization facilities. The framework shall be applicable to legal texts in different jurisdictions and over different legal/corporate documents. This necessitates that users should be able to customize the framework. Customization can be performed using, for example, a domain-specific language developed via the Eclipse Modeling Framework (EMF, http://www.eclipse.org/modeling/emf/). Such a domain-specific language will enable users to specify in an intuitive way, for example, the structure of legal texts, and the rules and heuristics for detecting semantic metadata items.

R6: Versatile NLP engine. The framework shall build on a versatile workbench for performing the NLP tasks. A potential candidate is GATE (http://gate.ac.uk) that we have been using in a variety of research projects in the past several years. A key advantage of GATE is that it brings together a large collection of mature NLP technologies and provides a unified mechanism for integrating them through a generic annotation infrastructure. This characteristic of GATE makes it possible to experiment with several alternative solutions. A second important advantage of GATE is that it has an "embedded" mode, allowing it to be easily integrated into different back-end services and workflows.

R7: Standardized markup format for legal metadata. Interoperability and seamless metadata exchange are important considerations to take into account. These factors necessitate a harmonized markup schema for encoding the extracted metadata. One should rely on standards and recommendations such as RDF (https://www.w3.org/standards/techs/rdf), RuleML (http://ruleml.org), LegalRuleML [17], or Akoma Ntoso (http://www.akomantoso.org) for metadata representation. The exact choice as to which schema(s) are most appropriate depends on contextual factors. The customization requirement (R5) therefore needs to encompass the tailoring and customization of the output markup language.

In our previous work, we used Akoma Ntoso for representing structural metadata. Akoma Ntoso is an XML markup schema for describing legal resources of various types, for example, laws, regulations and court decisions. In addition, we used an intuitive resource identification mechanism, ELI, for referencing and navigating resources. ELI (European Legislation Identifier, http://eur-lex.europa.eu/eli-register/about.html) is a European Union initiative aimed at providing a unified legal referencing mechanism. In particular, ELI defines a labeling framework based on a customizable template to enable the definition of universal resource names that are independent of the countries and legal jurisdictions to which the texts belong. ELI has been already adopted by several European governments and institutions, including the Government of Luxembourg for their legislative texts.

For semantic metadata, further investigation is required. Generic RDF triples and LegalRuleML are promising candidates for investigation.

C. Mapping Between Requirements and Desired Qualities

In Table I, we show how the different requirements of the framework (R1-R7) contribute to the desired qualities (Q1-Q5) discussed earlier.

R1 through R4 constitute the core functions, collectively satisfying the framework's effectiveness quality (Q1).

R4 further contributes to usability (Q4) by providing a UI for metadata editing and manipulation.

Customization facilities (R5) are critical not only for satisfying customizability (Q2) but also Q3 through Q5. Without models, adjusting the code to meet the structural characteristics of large collections of complex legal texts would not be scalable from a development effort standpoint. Models are also key to ensuring usability (Q4), due to the abstraction it provides in expressing the characteristics of legal texts from various countries and jurisdictions. Finally, models are important for integrability (Q5). More specifically, model-

to-text transformation technologies provide the flexibility to generate different implementation code targeted at different legal and regulatory contexts.

A versatile NLP engine (R6) provides a wide spectrum of tailorable modules (Q2). In addition, major NLP toolkits are highly optimized and scalable from a computational standpoint (Q3), and can be integrated into different workflows with relative ease.

Finally, relying on standardized formats such as Akoma Ntoso and ELI (R7) facilitates customizability, usability and integrability (Q2, Q4 and Q5) as a natural by-product of using standardized and interchangeable XML conventions.

IV. Next Steps and Conclusion

LegalTech is spurring considerable innovation, with numerous startups, SMEs and large corporations embracing it in order to reduce the cost of compliance and to facilitate day-to-day operations within the legal profession. Legal metadata is an important enabler for the innovation that is taking place in the legal domain. The research conducted by the RELAW community has important overlaps with LegalTech, including on the topic of legal metadata. We therefore believe that RELAW research has the potential to play a more prominent role in technology transfer and commercialization activities related to LegalTech.

In this paper, we presented our findings about the market landscape for legal metadata and highlighted the need for a flexible and robust legal metadata extraction framework. Our work is part of an ongoing proof-of-market study on legal metadata [18]. Drawing on our current findings, our next step will be to improve, and where necessary, to reconceptualize our early legal metadata extraction prototype [1], so that the prototype will better fit with market needs and practical considerations. We have already made some headway in this direction. In particular, we have been able to successfully generate accurate structural metadata at large scales in collaboration with the Government of Luxembourg [19]. However, much work remains to be done in relation to semantic legal metadata; this is indeed where we will be focusing most of our attention in the future.

Acknowledgments. This work has been partially supported by SCL (Service central de législation) and Luxembourg's National Research Fund (FNR) under grants FNR/P10/03 and PoC16/11554296.

References

[1] N. Sannier, M. Adedjouma, M. Sabetzadeh, and L. Briand, "An automated framework for detection and resolution of cross references in legal texts," *REJ*, vol. 22, no. 2, 2017.

[2] R. Vogl, "The Coming of Age of Legal Technology," https://law.stanford.edu/2016/09/26/184188/, September 2016.

[3] B. Goodman and J. Harder, "Four areas of legal ripe for disruption by smart startups," http://www.lawtechnologytoday.org/2014/12/smart-startups/, December 2014.

[4] J. Maxwell, A. Antón, P. Swire, M. Riaz, and C. McCraw, "A legal cross-references taxonomy for reasoning about compliance requirements," *REJ*, vol. 17, no. 2, 2012.

[5] T. Breaux, "Legal requirements acquisition for the specification of legally compliant information systems," Ph.D. dissertation, North Carolina State University, 2009.

[6] N. Zeni, N. Kiyavitskaya, L. Mich, J. R. Cordy, and J. Mylopoulos, "GaiusT: supporting the extraction of rights and obligations for regulatory compliance," *REJ*, vol. 20, no. 1, 2015.

[7] S. Ghanavati, L. Humphreys, G. Boella, L. D. Caro, L. Robaldo, and L. W. N. van der Torre, "Compliance with multiple regulations," in *ER'14*, 2014.

[8] "LIME - The Language Independent Markup Editor," http://lime.cirsfid.unibo.it.

[9] "LEOS - Open Source software for editing legislation," https://joinup.ec.europa.eu/software/leos/description.

[10] A. Phillips, R. Godfrey, C. Steuart, and C. Brown, *E-Discovery: An Introduction to Digital Evidence*, 1st ed. Delmar Cengage Learning, 2013.

[11] M. Hamdaqa and A. Hamou-Lhadj, "An approach based on citation analysis to support effective handling of regulatory compliance," *Future Generation Computer Systems*, vol. 27, no. 4, 2011.

[12] S. Ghanavati, D. Amyot, and A. Rifaut, "Legal goal-oriented requirement language (legal GRL) for modeling regulations," in *MiSE'14*, 2014.

[13] N. Zeni, E. Seid, P. Engiel, S. Ingolfo, and J. Mylopoulos, "Building large models of law with NómosT," in *ER'16*, 2016.

[14] G. Boella, L. D. Caro, L. Humphreys, L. Robaldo, P. Rossi, and L. van der Torre, "Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law," *Artif. Intell. Law*, vol. 24, no. 3, 2016.

[15] T. Breaux and A. Antón, "Analyzing regulatory rules for privacy and security requirements," *IEEE TSE*, vol. 34, no. 1, 2008.

[16] N. Sannier, M. Adedjouma, M. Sabetzadeh, and L. C. Briand, "Automated classification of legal cross references based on semantic intent," in *REFSQ'16*, 2016.

[17] T. Athan, H. Boley, G. Governatori, M. Palmirani, A. Paschke, and A. Z. Wyner, "OASIS legalruleml," in *ICAIL'13*, 2013.

[18] University of Luxembourg, "Open and transparent access to legal data," https://goo.gl/vpQCBD, May 2017, (Press Release).

[19] N. Sannier, M. Adedjouma, M. Sabetzadeh, L. Briand, J. Dann, M. Hisette, and P. Thill, "Legal Markup Generation in the Large: An Experience Report," in *RE'17*, 2017, to appear.