THE CASE FOR LINKING WORLD LAW DATA

Sergio Puig[*] and Enric G. Torrents[**]

[*] Stanford University
SLS professor and fellow at Codex
Legal Informatics Department
spuig@stanford.edu

[**] Autonomous University of Barcelona
Associate Researcher at the Institute of
Law and Technology
enricgarcia@uoc.edu

A revised, edited and updated version of this draft will appear in
2, JOURNAL OF OPEN ACCESS TO LAW (2014, Forthcoming)

— E PLURIBUS UNUM —

ABSTRACT

The present paper advocates for the creation of a federated, hybrid database in the cloud, integrating law data from all available public sources in one single open access system - adding, in the process, relevant meta-data to the indexed documents, including the identification of social and semantic entities and the relationships between them, using linked open data techniques and standards such as RDF. Examples of potential benefits and applications of this approach are also provided, including, among others, experiences from of our previous research, in which data integration, graph databases and social and semantic networks analysis were used to identify power relations, litigation dynamics and cross-references patterns both intra and inter-institutionally, covering most of the World international economic courts[1].

Author Keywords:

Linked Open Data, Semantic Web, Federated Databases, Data Integration, Data Analysis, Open Science Data Cloud, Neocodex

2012 ACM Classification Keywords:

• **Information systems~Information integration • Information systems~Deduplication • Information systems~Extraction, transformation and loading • Information systems~Wrappers (data mining) • Information systems~Entity resolution • Information systems~Federated databases • Information systems~Data cleaning** • *Information systems~Network data models • Information systems~Data exchange • Information systems~Mediators and data integration* • Information systems~Data warehouses • **Applied computing~Law** • *Social and professional topics~Database protection laws*

INTRODUCTION

As researchers experienced in the analysis of investor-state cases, trade disputes and the international arbitration field in general, we are well aware of the problems faced during the process of linking law data sources, and the reasons why, in some instances, law data is kept under lock and key.

We are not proposing absolute solutions to inconveniences such as poor data quality (unfitness for use), lack of unifying conventions, or the fact that any given system aiming to cover multiple jurisdictions soon faces challenges derived from the need to integrate materials in different languages, nor to more pressing issues such as lack of transparency and access to those sources that should be already public, but presenting a potential working solution, mostly automated, fit for purpose, reusable, referenceable, repurposeable and open to improvement through coordinated cooperation and collaboration of many, to integrate public legal data sources available on-line, include relevant meta-data to enhance indexing and categorization (thus also search and retrieving), while extracting and linking entities to create a higher layer of information and context, published in linked open data (LOD) standards.

Such is the scope of our current endeavor, Neocodex, a project backed by the Open Science Data Cloud[2] involving a growing network of researchers from several institutions, aimed to develop open source technology for integrating, analyzing and making available information from all international courts and national jurisdictions, including the mostly automated processing, analysis and

[1] Puig, Sergio "Social Capital in the Arbitration Market" *25 European Journal Of International Law* (2014, Forthcoming)

[2] http://www.opensciencedatacloud.org/projects/

visualization of social networks (neutrals, litigants, and other entities), semantic networks (citations, case-law contents, legal knowledge), and the publication of corpus collections with added meta-data.

THE POWER OF THE SEMANTIC WEB

LOD is, as Sir Tim Berners-Lee -creator of the Word Wide Web-, put it, "the semantic web done right". It is, thus, an essential element in the development and future of the Internet. As such, it has already become a powerful legal research and practice resource[3], whose growing capabilities are increased with every new addition of data-sets and information sources.

Projects like Matsu (in collaboration with NASA) and Bionimbus (with the institute for Genomics and Systems Biology), both of them also backed by the Open Science Data Cloud, provide excellent examples of how enabling for open access to relevant data can multiply exponentially the collective effectiveness of entire fields, pushing forward the limits of knowledge and human endeavor.

Each new digital cloud of well-curated information, when published in LOD standards[4], becomes part of a brewing, global storm called to reshape the information systems on which our society depends, and is based upon.

Current efforts by the European Union to integrate jurisdictional information from its Member States are illustrative of the need and benefits of linking data -specifically, law data- to foster integration, social and economic progress:

"Given the disparities between Member States' legal data at regional, national and European Union (EU) level, it is necessary to ensure that citizens have easy and efficient access to information on national and European legislation. The European Legislation Identifier (ELI) enables simple and fast access to this information, with a view to establishing the common area of freedom, security and justice." [5]

What we are advocating, and working for, though, is pursuing this kind of integration effort in a global scale, inter and intra-jurisdictionally -reaching for the maximum impact and benefits that LOD has to bring-. The technologies to do so exist, and are readily available to be adapted and implemented at will.

MAPPING AND RANKING THE SOURCES

Such ambition implies fostering and promoting the publication of law data from all jurisdictional courts, and the implementation of LOD standards from the source. Until this goal is reached, there are two main tasks to be performed. The first one is the publication of a report keeping track of the jurisdictions status by country, including metrics such as: degree of progress in the digitalization process, openness of the licenses used for publishing, and quality of the published data, inter alia, as described in the present section. The second one -the conversion to LOD standards of data from non-LOD sources, and the integration of all information into a single database-, is covered in subsequent ones.

An initial version of the report[6] has been published online using a CKAN based platform, a flagship project of the Open Knowledge Foundation[7]. This online version not only exposes the situation of law data sources to inform and educate governments, institutions and civil society -enabling for further actions to foster transparency and keep public bodies to account, a pursue essential on itself-, but also gives direct access to the sources' datasets and databases, where available. Ranking of the sources have been performed following a five stars scheme proposed by Sir Tim Berners-Lee:

☆	Data is available online
☆☆	Data is machine-readable
☆☆☆	Non-proprietary formats are used
☆☆☆☆	RDF[8] standards are implemented
☆☆☆☆☆	Data is linked to provide context

As well as specific law-related metrics.

CURATING RESOURCE IDENTIFIERS

[3] Berners-Lee, Tim. "The year open data went worldwide" *TED University*, February 2010

[4] Heath, Tom and Christian Bizer. "Linked Data: Evolving the Web into a Global Data Space" *Synthesis Lectures on the Semantic Web: Theory and Technology,* 1-136, Morgan & Claypool, 2011

[5] "Council conclusions inviting the introduction of the European Legislation Identifier (ELI)" *EU Official Journal C 325* (2012).

[6] http://neocodex.weboflaw.com/global-report/

[7] http://okfn.org/projects/

[8] RDF stands for Resource Description Framework, a W3C specification for the semantic web (web 3.0). http://www.w3.org/TR/rdf-primer/

The fourth stage in the data quality pyramid involves using unified resource identifiers (URI), so that users and agents can point at individual elements inside any given information repository. The creation of an URI scheme[9] and naming structure -curating a canonical list of entity identifiers, including a vocabulary list and an ontologies directory- is, thus, unavoidable to enable the process of linking global law.

The above is achieved either by determining the more suitable identifiers to use -either by adopting those used by the source (if any) or already in place and widely accepted by the community (i.e. ELI for the EU cases, GeoNames for geographical entities, DBpedia for Wikipedia entries, etcetera)-, or by assigning new identifiers to all entities: neutrals (judges, arbitrators), litigants (states, corporations, individuals), courts and centers, cases, documents, legal concepts, inter alia.

Given the lack of LOD standards implementation in most instances, a mixed approach consisting mostly of new identifiers is proving to be the most convenient solution. Further information on URI schemes and the status of the canonical list developed to cover all law resources can be found as an annex to the aforementioned report on law data sources.

DATA MINING AND PURIFICATION

The process of integration, by definition a permanently open-ended task, supposes the creation of, at least, a customized scrapping bot for each court allowing for the re-utilization of its information, and -in most instances- several bots per institution, developed to perform their objectives in the context of each court's publication habits and conventions.

This step not only allows for raw data gathering in a regular basis, but for entity extraction: identifying named entities present in the collected documents, turning such information into relations with other documents (weaving of networks), and meta-data used for indexing and categorization. Neocodex counts with the support of, among others, Outwit Technologies[10], as well as open source solutions such as Nomenklatura[11], and a growing corpus of own code to overcome the inherent difficulties of this process.

Pulling the data from the source to its new form, re-published as LOD, often involves format conversion and processing of natural language, non-structured, non-machine readable data. The further down a source is in the data ranking presented before, the more complex the data mining and purification steps are.

As stated, one of the main priorities is to ensure the implementation of best practices at source. For this purpose, technologies developed during the course of this project will be freely shared with interested courts to facilitate the publication of LOD at the point of origin. This, in turn, will make it possible to go far beyond in the integration process.[12] Unfortunately, at this stage the project hasn't yet reached any agreement of direct cooperation with any national or international court or center.

USE CASE: INTERNATIONAL COURTS

Until now, most of the research derived from the usage of this incipient LOD cloud of legal information has focused in the study of inter-jurisdictional social networks in international economic courts and arbitrations centers, such as ICSID, PCA, ITLOS, the Iran-United States Claims Tribunal and WTO, as well as other international courts such as ECHR, ICC, ICJ-CJI, ECJ-CFI, OHADA, the African Court on Human and Peoples' Rights, the Inter-American Court Of Human Rights, the Court of Justice of The Andean Community, and administrative tribunals such as those of the Inter-American Development Bank, International Labor Organization, International Monetary Fund, World Bank, OECD, Organization of American States, and India Central Administrative Tribunal, inter alia.

We have studied the dynamics of legal knowledge transmission between arbitration centers and international economic courts by identifying cross-reference patterns, legal concepts usage, and the role of bridge individuals, as well as the oftentimes imbalanced distribution of appointments that leads to the concentration of an elevated number of decisions in the hands of a few power-brokers, setting, in turn, precedent for further decisions by other neutrals. By doing so we have identified and analyzed the activity and role of highly

[9] http://www.w3.org/wiki/UriSchemes

[10] http://www.outwit.com/

[11] http://nomenklatura.okfnlabs.org/

[12] Bechhofer, Sean, et al. "Why Linked Data is Not Enough for Scientists" *Future Generation Computer Systems* (2011).

active individuals present in more than five courts and centers, among more than ten thousand players. These are, among others: Francisco Orrego Vicuña, Stephen M. Schwebel, Jan Paulsson, Charles N. Brower, James R. Crawford, Yves L. Fortier, Karl-Heinz Böckstiegel and Florentino Feliciano, as well as over one-hundred fifty neutrals active in at least two international courts or tribunals.

These ongoing analysis, presented online in the form of interactive network visualizations[13], constitute a natural extension to previous research on the social capital in the arbitration market, published at the European Journal of International Law. Similar studies are under way to discover hidden patterns in human right courts and cases, and other applications.

CONCLUSION

Our research, as many others, would had been impossible without being able to study the whole, rather than just the sum of its parts. We are convinced that, in an ever more interconnected World, being able to explore all law data at once, enabling not only for human review but also for computational analysis of all information, plus having the capability to further contextualize the data by connecting it with other LOD resources (government data, economic indicators, etcetera), is becoming a necessity more than a luxury. A comprehensive LOD resource devoted to law will become, once fully functional, an indispensable tool to understand and improve the legal systems.

We invite any individual and organization to join in and participate in this open endeavor, to shape together this project, Neocodex, aspiring to replicate the impact that Justinian's Corpus Juris Civilis, the original Codex, had in the legal systems of the Early Middle Ages.

AKNOWLEDGEMENTS

Special thanks for their remarks, support and contributions to Walt Wells (Open Science Data Cloud), Sébastien Heymann (Gephi, Linkurious), Friedrich Lindenberg (Open Knowledge Foundation, Mozilla, Der Spiegel), Clemens Wass (OpenLaws), Hakim Djoumi (Amadeus), and Phil Archer (W3 Corsortium).

[13] http://weboflaw.com/visualizations.html