Linked data in the legal domain

T. Agnoloni, E. Francesconi, M. T. Sagri, D. Tiscornia[1]

Abstract Despite the increasing coverage, the current scenario in accessing legal information shows a proliferation of different points of access delivering heterogeneous data, organized according to different criteria and formats, provided either by public producers or private publishers, from different countries and in several languages. Data are often closed in inaccessible databases and without any chance to establish an interoperability layer among the different source of information, to allow reuse and interconnection. The current momentum gained by Open Government Data and Linked Open Data initiatives offers a unique opportunity to foster an effective free and open access to legal information. The paper introduces a project for creating Legal Linked Data, currently carried on at ITTIG and aimed at providing the methodological premises and the technical building blocks on which a new comprehensive information service for legal experts and innovative applications for citizens could be designed. Ittig will re-use and integrate in a global system all the experiences, tools and data produced in its long research experience, the core of the project being the DoGi (Dottrina Giuridica, legal literature) database, created and distributed by Ittig since 1969.

Introduction

Nowadays the amount of available unstructured (or poorly structured) legal information made available as part of public accessibility projects by governments, free access initiatives and web portals has reached an unprecedented coverage and will probably keep growing as the web expands. However, a coherent logical organization of the available material and an effective accessibility of legal sources in terms of re-usability, search functionalities and completeness of coverage around a legal topic has still to come.

Legal information are spread out in a proliferation of local sites, organized by subjective criteria where data are stored and represented in different formats,

[1] Institute of Legal Information Theory and Techniques of the Italian National Research Council (ITTIG -CNR)
{agnoloni,francesconi,sagri,tiscornia}@ittig.cnr.it

often closed in proprietary databases, so that it is quite impossible to define an interoperability layer among the different sources of information. Most of the effort is left to the user to query, collect and integrate the information to overcome the technical barriers.

Many initiatives in the legal informatics research community have been promoted to deal with such issues. In particular semantic enhancement of legal information [1] and national and international initiatives on legal documents standardization. However, despite the relevance of such researches the *top down* approach they underly has so far been an obstacle to the actual implementation on a large scale of the envisaged solutions as they would require a wide coordination and economic effort of the involved actors to adopt the proposed standards with little immediate benefits.

The current momentum gained by Open Government Data and Linked Open Data initiatives around the world offers a unique opportunity to foster an effective free and open access to legal information according to the linked open data principles joined by the complementary standardization initiatives brought on by the legal information community.

Linked Open data

One of the keys to success of the Linked Open Data (LOD) initiative is that, while based on exactly the same technological stack and principles of the Semantic Web, it proposes a pragmatic bottom up approach to refine and enrich the web resources sketching an immediately viable path towards the incremental realization of the semantic web vision. The first step towards this change of perspective is the exhortation to data owners to *"liberate the data"* or *"raw data now" i.e.* to put the raw data available and accessible in whatever open format, first necessary condition to implement the overlying knowledge organization layers.

By adopting the linked data best practices of publication [2] (first of all on existing databases), we put the basis for a growth free of technical barriers of enhanced content which will constitute the basic building blocks to be later reused to feed more advanced knowledge intensive systems. As cited in [3] one of the barriers to accessible law is that:"*To a worryingly large extent, statutory law is not practically accessible today, even to the courts whose constitutional duty it is to interpret and enforce it. There are four principal reasons. … First, the majority of legislation is secondary legislation…. Secondly, the volume of legislation has increased very greatly over the last 40 years … Thirdly, on many subjects the legislation cannot be found in a single place, but in a patchwork of primary and* *secondary legislation. … Fourthly, there is no comprehensive statute law database with hyper links which would enable an intelligent person, by using a search engine, to find out all the legislation on a particular topic."* The linked data model seem to provide the technological answer to such a question of fragmentation of legal sources by providing the infrastructure to seamlessly collect in a single place document fragments from distributed sources.

Open Data and Legal Standars

The application of the linked open data best practices of publication in the legal information domain finds fertile ground in mature initiatives for the definition of standard formats for the identification and representation of legal documents on the web. Here we mention the most relevant.

Legal sources identification

URN:lex is a proposed Internet standard for legal document identifiers[2]. The URN:LEX namespace aims to facilitate the process of creating URIs for legal sources independent of a document's online availability, location, and access mode. "Sources of law" include any legal document within the domain of legislation (including bills), case law and administrative acts or regulations. This identifier will be used as a way to represent the references (and more generally, any type of relation) among the various sources of law[3].

Legal document structure representation

Initiatives on adoption of XML standards for the representation of legislative document structures and metadata have been brought on both at national and international level in different countries in recent years. They all basically aim to provide a Open XML interchange format for legal and legislative resources thus conforming to requirement of opennes in LOD. To cite the most successful, XML.gov in the U.S., Crown XML Schema in the U.K. provide the most rich and complete datasets made available by governments in open XML. Other initiatives in European countries, like NIR (NormeInRete) standard in Italy or Metalex in the

[2] http://tools.ietf.org/html/draft-spinosa-urn-lex-01

[3] A different but compatible approach have been proposed and implemented in the definition of the URI identification schema of British Legislation in one of the most advanced initiatives on publishing legislative data in open format in the portal www.legislation.gov.uk described in [4]. Here a Persistent HTTP URI scheme have been used following the design principles conforming to the recommendations of W3C for the design of URIs for the semantic web [5];

Netherlands have also lead to further development for a panafrican standard (AkomaNtoso) and to the international initiative of Metalex/CEN[4] global interchange standard of legal sources.

Metadata scheme

In legislation.gov.uk a sophisticated metadata model — incorporating FRBR[5], the CEN MetaLex vocabulary, Dublin Core Terms[6], and the Crown Legislation Markup Language — enabling advanced version control and output of descriptive metadata have been adopted providing also all the metadata able to implement a point in time legislative system. In general a very minimal metadata set would consist of a title, an effective date, the name of the issuing body and some sort of permanent identifier (for example conforming to the urn:lex specification). Beyond that, one might add more dates (like date of efficacy, date of publication, dates of stages in the process of drafting and approval), compact descriptions of the legislation and its intended effects, a representation of the legislative process, responsible agencies and organizations, and so on[7]. The more descriptive metadata are attached to the document (or even better, at subparts of it), the more meaningful interconnections among different documents can be established.

In this respect one of the advice in the reference guidelines for Putting Government Data online [6] is that “There are two philosophies to putting data on the web. The top-down one is to make a corporate or national plan, by getting committees together of all the interested parties, and make a consistent set of terms (ontology) into which everything fits. This in fact takes so long it is often never finished[...]. The other method experience recommends is to do it bottom up. A top-level mandate is extremely valuable, but grass-roots action is essential. Put the data up where it is: join it together later. Do not wait until you have a complete schema or ontology to publish data.”

Legacy Database Schema as source of Metadata

Sticking to such philosophy a fundamental source of metadata are existing legacy relational database structures and schemas. Once they are transformed in an RDF schema and made openly accessible on the web according to the same Linked Data principles, they constitute a fundamental and immediately available KOS (Knowledge Organization System) to the underlying instances of the

[4] *www.metalex.eu*

[5] *http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records*

[6] *http://www.dublincore.org/documents/dcmi-terms/*

[7] Suggested metadata practices for legislation and regulations *http://topics.law.cornell.edu/wiki/lexcraft/suggested_metadata_practices_for_legislation_and_regulations*

database. Indeed there are a number of open source tools for putting relational databases up as Linked Data e.g. D2RQ platform[8] and Triplify[9] being two. These are able to analyse the schema of an existing database to create a default mapping file explaining how the database structure actually represents things and their relationships. The result of the mapping is a translation of the relational database into and RDF Schema with entries of the DB exposed as RDF instances on the web of data, thus making the original database structure an immediately available source of metadata and interlinked relations surrounding the documentary units.

An Italian project

Despite the increasing number of good practises [7], a comprehensive information systems for lawyers, able to cope with linguistic and conceptual barriers still doesn't exist. What is missing is a basic overall framework in which heterogeneous legal information can be recognized and interconnected even if by means of shallow conceptual links. Our aim is to start, from these premises, the incremental construction of a "*Legal Data Cloud*": an example, relating to some of the Italian and European Union relevant legal datasets, is sketched below.

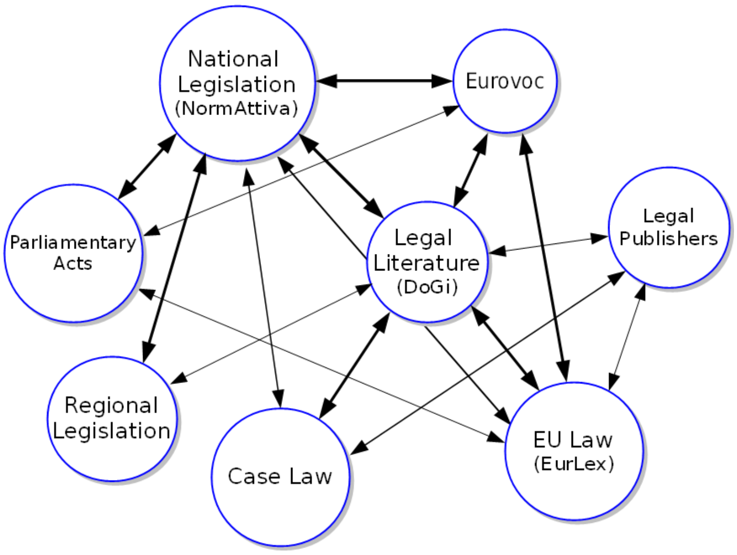

Fig. 1. A possible interconnection of relevant Legal Datasets in a "Legal Data Cloud".

[8] *http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/*

[9] *http://triplify.org/*

The legal data cloud should link legal knowledge units from proposal and debate (Parliamentary acts, Regional councils) to publication (National Legislation portals like NormAttiva in Italy, Eur-Lex Access to European Law, Regional legislation datasets), to interpretation (legal literature) and case law.

The interconnection *glue* would be, at a first stage, built on the RDF translation of the metadata scheme of the existing databases, further connected by additional mapping RDF statements. Furthermore, several datasets, could be semantically interconnected through EuroVoc, the EU's multilingual thesaurus, recently converted in SKOS/RDF, also enabling cross-lingual retrieval.

Up to now, and despite several no-profit initiatives[10], the availability of data published by Italian Public Administrations in open format is very low, with few brilliant exceptions at regional level, like the Piedmont region portal (*www.dati.piemonte.it*). In the legal domain, altough the small amount of data available in open format, ITTIG has started working on a concrete application: the aim is to provide the methodological premises and the technical building blocks on which a new comprehensive information service for legal experts and innovative applications for citizens could be designed. In the methodological setting, technical solutions are addressed to :

- **re-using** all information (metadata) and domain specific standards already available. Thist requirement has been met since ITTIG is a legal information provider by itself. We intend to collect and integrate in this initiative all the experiences[11], tools[12] and data that ITTIG has produced in its long research experience, the core of the project being the DoGi database, produced and distributed by Ittig since 1969. DoGi is, in the Italian national context, one of the most relevant sources for online research on legal literature. It provides abstracts of articles published in more than 250 Italian periodicals. Documents in the data base (currently 350.000, with an annual increase of 14.000 DoGi documents) refer to periodical articles, comment on cases, commentaries to statutes, conference papers, book reviews. The reason why we focus on Dogi as the basic block of knowledge is the rich set of metadata added to each document, including bibliographic records, abstracts, classification codes, a flat list of key words and references to legislative (EU[13] and national[14]) sources and to case law. In addition, the data format used in DoGi (enabling XML export and with metadata

[10] Open Knowledge Foundation Italia (*it.okfn.org/)*, Openpolis (*www.openpolis.it/)* and *www.spaghettiopendata.org/.*

[11] see, for instance *www.dalosproject.eu*

[12] *www.xmleges.org*

[13] Retrieval of the full-text of European legal texts through links to CELEX database)

[14] Retrieval of the full-text of Italian legislation (from 1946) through links to the freely-accessible legislation portal *www.normattiva.it*

schema based on Dublin Core) is the best option for the conversion in open format.

At this early stage of the project, the main effort is devoted to the transformation of the DoGi in a rich repository of linked data; the non ambigous semantics of normative references enables the automatic linking of legal literature to legislation already available through national and European sites, to bibliographic catalogs, and to available case law, as for instance the Italian Constitutional Court's information system[15]. In the next step the 'legal cloud' will be enriched by linking data to a selection of Italian case-law[16] and to the (few) legal data opened by the Italian public producers, as for instance, the catalog of public organizations and services[17].

- **integrating** semantic resources into a shared conceptual layer and creating technical pre-conditions for collaborative and coherent semantic enrichment. A specific platform for thesaurus mapping and ontology building is currently under development, based on the VocBench tool and on the analysis of existing commercial and free tools[18]. It will enable the re-use of available semantic resources[19], thus favoring collaborative data integration to external resources[20] and conceptual linking.

- building **citizen-oriented applications.** The main idea is to provide non expert users with a simplified vision of legal texts. by implementing a *Provision Model* [8], that offers a functional vision of the legal system as a set of provisions. The automatic (or semi-automatic) bottom-up detection of provisions from text requires tools able to classify provisions, as well as to extract their attributes (e.g. the Bearer of a Duty). Machine learning and NLP tools have been used to implement such automatic processing [9]. The model has been recently represented using RDF/OWL standards and specific axioms, derived from fundamental normative relations, like the equivalence between Duties and Rights in their implicit and explicit views, have been introduced using OWL-DL expressivity [10]. A Provisions RDF triplestore can be queried to retrieve specific type of provisions, involving specific entities, using SPARQL. Such a query will be able to retrieve all the Rights contained in the collection identified by a specific URI and, in case the inferred model is queried, all the inferred provisions are

[15] *http://www.cortecostituzionale.it/*

[16] *www.caselex.org*

[17] *www.lineaamica.gov.it/rubricapa*

[18] Poolparty (*poolparty.punkt.at*), Voc-bench (*www.fao.org/agrovoc)* Ontowiki *(ontowiki.net)*

[19] the LOIS multiligual legal WordNet, the CLO core ontology (*wiki.loa-cnr.it/index.php/LoaWiki:CLO*), the Dalos ontology (*www.dalosproject.eu/).*

[20] among others, the Legal Taxonomy Syllabus (www.eulawtaxonomy.org).

retrieved, for instance: either annotated as Rights of the Consumer or those implicitly deduced, namely the Duties where the Consumer is a counterpart.

Conclusions

The rapid growth of open government data initiatives around the world, based on the technical recommendations of the linked open data movement, is a unique occasion to foster the publication of legal datasets in open formats.

Thanks to a layered standards infrastructure, Linked Data best practices can be transposed as is to legal data publication in order to reach a critical mass of legal information available in linked data format to put in practice the vision of innovative legal semantic web applications and services built on top of a "Legal Data Cloud". In this view we propose the reuse and interconnection of existing tools, standards, formats and legal datasets developed by ITTIG over the last decade to bootstrap the process of enhanced legal open data publication in the Italian context. Reuse of existing legal metadata scheme from available legal datasets on the web can be a starting point to bootstrap an iterative process of refinement towards a fully interconnected open legal semantic web.

References

1. Casellas, N. (2010), Semantic Enhancement of Legal Information… Are We Up for the Challenge?, available at http://blog.law.cornell.edu/voxpop/2010/02/15/semantic-enhancement-of-legal-informatiom...-are-we-up-for-the-challenge/
2. Heath, T. and Bizer, C. (2011), Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.
3. Holmes, N. (2011), Accessible Law , available at http://blog.law.cornell.edu/voxpop/2011/02/15/accessible-law/
4. Sheridan, J. (2010), legislation.gov.uk, available at http://blog.law.cornell.edu/voxpop/2010/08/15/legislationgovuk/
5. Sauermann, L. And Cyganiak R. (2008), Cool URIs for the Semantic Web, W3C Interest Group Note, available at http://www.w3.org/TR/cooluris/
6. Berners-Lee, T. (2009), Putting government data online, available at http://www.w3.org/DesignIssues/GovData.html
7. Boella G,, Humphreys L., Martin M., Rossi P. and Van Der Torre L., Eunomos, a legal document and knowledge management system to buildlegal services, presented at AICOL 2011, August 16-17, Frankfurt am Main
8. Biagioli C. and Grossi D. (2008), Formal Aspects of Legislative Meta-Drafting, in: Francesconi E., Sartor G. and Tiscornia g. (eds.), "Legal Knowledge and Information Systems - JURIX 2008, 192-201, Amsterdam, IOS Press.
9. Francesconi E. (2011), A Learning Approach for Knowledge Acquisition in the Legal Domain, in: Sartor g. Casanovas P., Biasiotti m.A. and Fernández-Barrera M. (eds.), "Approaches to Legal Ontologies, Theories, Domains, Methodologies", Berlin, Springer.
10. Agnoloni T. and Francesconi E. (2011) *Modelling Semantic Profiles in Legislative Documents for Enhanced Norm Accessibility,* Proceedings of ICAIL 2011, N.Y., ACM Press.