Model and ontology based conceptual searching in legislative XML collections

Carlo Biagioli
Ittig-Cnr
carlo.biagioli@ittig.cnr.it

Fabrizio Turchi
Ittig-Cnr
fabrizio.turchi@ittg.cnr.it

abstract

The migration of legislative archives into Internet, the imminent "semantic web" and the expectations generated from such processes make it necessary and possible, thanks to the technologies now available, to proceed to the review of information systems and data. The access of the citizen to the legislative data will be in fact not only virtual but effective only if the informative systems will be adapted to such not specialized user. It will be necessary to modify the systems and the data: appropriate interfaces will make it easier and in some way guide the access of common user. The data also should be transparent and enriched with meta data, able to point out and make it comprehensible the meaningful aspects, also according to various user profiles. In this paper we present a model able to describe the illocutionary profile of legislative texts through meta data and an application for conceptual retrieval, based both on the model and on domain concepts ontologies.

1. Formal and functional text patterns

Documentary systems are aimed at identifying, in large-sized collections, the documents meeting the user’s requirements – whether by subject-matter or not. To that end, they are provided with all the information allowing the individual documents to be distinguished as well as with summary guidelines on the basic subject. After the relevant documents have been found, the search is over: fruition of the documents is a step falling outside the scope of documentary systems. Nowadays archives including fully on-line documents are increasingly common: in this case, fruition is simultaneous with retrieval of the information.

Furthermore legal documents show peculiar features and are quite different from other documents in terms of their fruition. Indeed, it can be argued that the reason underlying a considerable portion of documentary researches on law collections is related to the need for identifying rules rather than law texts as such – which are therefore, at the very most, an intermediate target. To simplify things considerably, one might consider a rule to be the outcome for an operation in which pieces of various texts are put together based on specific criteria – just like in a jigsaw puzzle. If this is the case, an effective documentary system should facilitate retrieval of the significant pieces required for building up the rule(s) sought. Therefore, information retrieval should be followed by information extraction. In order to facilitate the information extraction activity, it will be necessary, in the first place, to identify, highlight and describe the individual pieces as included in the relevant texts. In the model presented here (Biagioli, 1997), the significant pieces are regarded as provisions and their necessary components are their arguments, while the conditionals elements and elements regarding time and space (acting on both provisions and arguments) are called operators. Although no linear relationship can be found between functional and formal profiles in a text, or between illocutionary and rhetoric structures, as brilliantly Branting called them (Branting, 1997), there is a reassuringly consistent correspondence
between text units (paragraphs) and meaning/functional units (provisions). Based on this finding, it can be reasonably argued that this is no casual correspondence: indeed, each provision is, from a general standpoint, the “meaning” of each basic (elementary) text unit. Knowledge of the basic unit-provision pattern is somewhat already included in legal texts – as shown by the titles of the individual articles and by those of the various parts in a law, i.e. by the meta-textual messages provided by lawmakers in order to facilitate the text understanding. This kind of knowledge will be used in order to identify, describe and, subsequently, search for the individual provisions. These authentic annotations basically show the regulatory “intentions” and/or the corresponding arguments.

To enhance effectiveness of this description, the information is re-phrased in a schematic, systematic fashion according to the “Functional Structure” model and are converted in metadata, possibly by including supplementary information as derived from the text of the individual units whenever this is considered necessary and the corresponding titles are not sufficiently explicit.

The final outcome may range – depending on the approach chosen – from re-wording of the original annotations up to a complete description entailing linguistic and legal construction of the individual pieces – without prejudice to the type of information conveyed.

In this case preference was given to adherence to the original text, by limiting interpretation and integration to the greatest possible extent. When the law maker, in the formulation phase of the new law, will provide accurate annotations with the aid of legislative drafting support software, these actions will not need any more.

2. Types of provisions

“Semantics” of a text is the result both of its atomic components (simple and complex terms, syntags) and of the aggregation of such components (simple and complex clauses, which might be likened to linguistic acts – i.e., to normative micro-acts) and of the text as a whole (the normative macro-act). Functional structure includes classes and types of provisions and arguments of provisions, as well as operators on provisions/arguments (conditions, time, space)). This classification has been derived from legislative techniques and practices, by taking also account – where necessary – of the best-known theories on norms (Hohfeld W.N. 1913).

A peculiar category includes the provisions related to the dynamics of a legal system (amendments), while another includes provisions on law (e.g. provisions laying down the scope of a law).

There are then the two major (and more concerned with an information system) categories related to substantive rules, which are the subject of normative theories, often referred to as constitutive rules and regulative rules, respectively based on the well-known distinction drawn by (J. Rawls, 1955): “justifying a practice and justifying a particular action falling under it”, where practice means “any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure”.

3. Arguments

Arguments are the main focus of the individual provisions: they are always to be found in a provision, indeed no provision could be regarded as such in the absence of its own argument (see, for instance, the definiendum of a definition-provision). In addition to the type of provision, they typically make up the contents of section/part titles. Arguments may be either single or multiple, reflecting the current styling of a provision, which facilitates their conversion into the descriptive formulas used in this model.

In case of complex arguments, to reach a detailed description, they can be also divided into subarguments. Arguments such as action/activity are inevitably more complex than subject/addressee arguments –which are atomic in nature. In order to increase effectiveness of their use, arguments will have to be synthetized into
more compact descriptions. Moreover they will be connected to outstanding descriptions, as dictionaries and ontologies, where terms are explained through connections to others terms and central concepts.

4. Operators

Provisions in the model proposed here are accompanied by conditions, widely intended as specifications of the central mining of the arguments or provisions they are connected to. The simplest way to highlight conditions consists in “marking up” the passage including them, thereby defining a significant area where word-based searches can be performed. A more synthetic approach can also be developed by identifying key-terms, connected by propositional operators, to summarise the conditions.

Besides the general conditions, there are some concerning time and space. Considered their relevance, those elements (time and space) will be deeply and formally described in the model.

5. Anchorage of the arguments to ontologies

Finally, as pointed out, it is previewed the connection of the arguments of the provisions to dictionaries and ontologies, in order to add the static-linguistic knowledge, to the dynamic-normative knowledge deriving from the model of the provisions. The model allows to evidence terms or sintagma that have a meaningful role in the provisions (e.g. addressee). Such terms often are defined from the same legislator in the definitional provisions or they are derivable from other legislative acts of the same domain. Where those explicit definitions have not been previewed, it will be possible to connect the terms to external descriptions, dictionaries and ontologies of domain.

At the end of the marking up process, a self explaining text will be available, able to inform the user on the essential content of each provision, both in its normative and thematic profile. Such knowledge, beyond facilitating the search of the norms and the consultation and understanding of the documents, will also support other processing tasks such as legal reasoning applications.

6. Meta-searching: an example

Let us suppose a user interested on italian provisions ruling personal data processing and a classic retrieval engine able to find several acts as temporary answer. Between them the following act will be available in the transient data base:


In a second step of the query the user can ask the system to extract from the previously founded documents, fragments marked up as obligation provisions. Several provisions of such kind will be showed.

This is the first result of a system based on analytical metadata derived from a model of provision: asking for provisions and not only for documents is provided by the system.

A next step in the search refining would be the delimitation of the founded obligation provisions to a sub-class regarding a particular ruled action. In the general provision model, the obligation provision has its own components: arguments as addressee, action and counterpartie; operators as condition, time, space.

Let us suppose the user is interested just in the obligations related to formal fulfiments. *MetaSearch* will give the user the chance to choose an external conceptual structured dictionary and navigating on it, to find a term useful in defining the kind of action he has in mind.
Let us suppose he will have a look in the tree of the dictionary focusing on a very general term as for instance “Legal act”.

![Legal act]

Figure 6.1: Conceptual dictionary/ontology

Supported by the conceptual dictionary, the user can choose for instance the “notification” term. The new answer will contain several obligation provisions, all of them regarding a special “action” defined by the concept chosen in the dictionary and by all the hyponyms terms-children in the dictionary-tree.

For example:

CHAPTER II OBLIGATIONS RELATING TO THE CONTROLLER

Article 7 (Notification )

1. A controller intending to process personal data falling within the scope of application of this Act shall have to notify the Garante thereof, exclusively in the cases and manner set out in the regulations as per Article 33(3), if the processing is liable to adversely affect the data subject's rights and freedoms on account of either the relevant mechanisms or the nature of the personal data (*).

Another chance is provided by metaSearch. Some kind of provisions are connected by a special relation: for instance an obligation can be linked to a penalty: in fact they can be viewed as parts of the same norm. The user can ask the system to find also penalties related to the obligation provisions previously found.

CHAPTER VIII PENALTIES

Article 34 (*)

(No other penalties are related to the obligation above.)

1. Whoever fails to promptly submit the notification required under Articles 7, 16(1) and 28 or provides incomplete information in a notification, in breach of his/her duties, shall be the subject of an administrative sanction entailing payment of an amount ranging between Lit 10 million (i.e., euro 5.164,6) and Lit 60 million (i.e., euro 30.987,4), as well as of the additional sanction consisting in publication of the relevant injunction/order.

7. Marking-up: an example

The “semantics” underlined by the model is expressed through the metadata. In other words the type of each provision and the meaning of its arguments will be translated into analytical metadata. Let us have a look to the XML-like marked up source data, to understand how the system reached those conclusions.

Marked up obligation provision:

CHAPTER II OBLIGATIONS RELATING TO THE CONTROLLER

Article 7 (Notification )

1.<Obligation> <addressee> A controller </addressee> <condition> intending to process personal data falling within the scope of application of this Act </condition> shall have </action><def>to notify </def> the Garante thereof, </action>……….<Obligation>

Marked up related penalty provision:

CHAPTER VIII PENALTIES

Article 34 (*)

1.<Obligation> <addressee> A controller </addressee> <condition> intending to process personal data falling within the scope of application of this Act </condition> shall have </action><def>to notify </def> the Garante thereof

Marked up related penalty provision:
1. Whoever fails to promptly submit the notification required under Articles 7, 16(1) and 28 or provides incomplete information in a notification, in breach of his/her duties, shall be the subject of an administrative sanction entailing payment of an amount ranging between Lit 10 million (i.e., euro 5,164,6) and Lit 60 million (i.e., euro 30,987,4), as well as of the additional sanction consisting in publication of the relevant injunction/order.

8. metaSearch: research software of provisions metadata

In order to extract knowledge from provisions metadata, defined in documents by XML tags, has been developed metaSearch, a software that allows to query the functional (illocutionary) profile of texts.

Documents are stored in collections made by sets of acts concernig a chosen juridical domain that can be queried as a whole or one by one single act.

The metaSearch software uses XML technologies and a database relational, metaCollection, constructed in asynchronous phase, containing metadata provisions and their arguments. The generation of the database carries out through a SAX parser that collects the metadata data of the provisions (view Figure 9.1). Such parser doesn't build any structure in memory to represent the XML document; this makes SAX fast and highly scalable, as the application builds exactly as little or as much in-memory structure as needed for its specific tasks (C.A.Jones, F.L.Drake Jr , 2001).

The metaSearch software is composed by various modules (view Figure 9.2), all based on a web interface and rely on the metaCollection database. A DOM parser is used to extract the full-text provisions and to activate the hypertext navigation. The use of DOM is particularly effective because for particularly complicated processing taks involving reasonably small XML documents, it's preferable to let the library build in-memory structure that represent the whole XML document and then traverse those structure. In the phase of information extraction we have also used the Xpath, a standard W3C, because some arguments' value of the provisions are expressed by its formalism. A simplified example is the following:

```xml
<paragraph id="art16-com3">
  <disposition>
  </disposition>
  <penalty>
    <action>paragraph[@id="art16-com2-let2"]</action>
  </penalty>
</paragraph>
```
where we emphasized the Xpath expressions, used to set the values of arguments.
In the future we could use a XPath query engine: to query the documents collections you could use XPath as defined by the W3C. This provides a reasonably flexible mechanism for querying documents by navigating and restricting the result tree that is returned.
Furthermore we could make available the software as XML-RPC API: it's a spec and a set of implementations that allow software running on disparate operating systems, running in different environments to make procedure calls over the Internet. It's remote procedure calling using HTTP as the transport and XML as the encoding. XML-RPC is designed to be as simple as possible, while allowing complex data structures to be transmitted, processed and returned.

9. metaSearch: software design

The metaSearch design can be represented by Figure 9.1 e 9.2 (asynchronous and interactive phases)
results: the final user will be able to choose to include also the terms in ontological relation with the chosen one so to widen the range of the results founded.

There are also other alternative as regard our solution: we could use a native XML search engine as Xindice from the Apache Xml Project, a database designed from the ground up to store XML data or what is more commonly referred to as a native XML database, Xindice currently supports XPath as a query language. In many applications XPath is only applied at the document level but in Xindice XPath queries can be executed at either the document level or the collection level.

10. Conclusions

Search for rules rather than texts

Formal description of a text results into defining formal documentary units – such as texts (of laws) and articles. Semantic model based description adds virtual documentary units of substantive character to information systems – in this case, normative provisions. The traditional search for legislative texts or articles will be coupled with the search by normative provisions.

The system will be required to yield units (articles, paragraphs, etc.) including specific types of provision to be found in various laws at the same time. This will allow identifying all the requirements, obligations, delegations, amendments, etc. as provided for by all the laws regulating a given subject matter.

With the help of conceptual descriptions of the relevant normative sector (domain-specific legal ontologies), each query will be better detailed and the requirements to be met by specific entities will be identified – by also taking advantage of the knowledge included in the arguments of the individual provisions.

Additionally, substantive documentary units can be a source of further information which is not included in documents but can be deduced from them based on the legal knowledge. For instance, existence of a rule concerning the claim with B as addressee and A as counterpart about the action X, can be deduced from the existence of a rule imposing an obligation on A addressee towards a B counterpart about X, based on logic relationships between deontic concepts. This type of inference can be activated so as to expand (at the user’s request) system answer. These rules are not laid down expressly by lawmakers for the sake of brevity – exactly because they can be deduced via the meta-rules of the normative system.

The metainformation envisaged here can be useful in many respects:

Dissemination and multilinguism

It can be reasonably argued that this system facilitates consultation of normative collections by non-expert users, who are probably the majority of Internet users.

Concept-based research can also be expected to enable access to information in multilingual version, as it is based on correspondence between concepts.

Diagnosis on provisions

This has to do with facilitating identification of critical sections in a text – i.e., provisions requiring further consideration by legal scholars in order to decide whether flaws or mistakes are present. There may even be conflicts between express provisions as shown by text annotations: this may be the case of non-homogeneous definitions of the same entity to be found in different legislative texts.

Monitoring for public administrative agencies

It will be easier for public administrative agencies to identify tasks and functions committed to the individual organs as resulting from all the relevant laws and regulations. This should allow assigning additional tasks based on adequate knowledge of the workload and therefore, of the “administrative feasibility” of the planned organisation.

Drafting laws

These same models may also be used as guidelines in drafting new laws. Drafters will be facilitated in their activity, as the required checks will be managed directly by the system and the relevant standards will be complied with thoroughly and in an automatic fashion.
11. REFERENCES


