STIA: Experience of Semantic Annotation in Jurisprudence Domain

Maria Teresa Pazienza, Noemi Scarpato, Armando Stellato,

ART Group, Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
{pazienza, scarpato, stellato} @info.uniroma2.it

Abstract. In this work we present STIA: a tool for semantic annotation in the Jurisprudence domain. The tool offers an easy interface to domain experts (lawyers, administrative, researchers,…) for annotating relationships of pertinence between portions of text from different laws covering similar topics/circumstances/events. These annotations both constitute a resource on their own (which can be used inside semantic search engines to easy retrieve related laws) as well as a precious feed for tools aiming at automatically extracting more of the above relationships.

Keywords. Semantic Annotation, Semantic Web, collaborative tagging

Introduction

Managing the information represented in large collections of documents is one of the main problems inside public administrations: the size of documental archives is in continuous growth and the peculiarities of the legal domain imposes searches which may span across decades and a plethora of information sources.

In the legal domain in particular, the spread of norms and laws containing explicit cross references or, even worst, overlapping over same/similar topics, events and situations has brought to various actions (in several states of Europe and in the EU itself) for legislative simplification: it is not only a matter of reducing the amount of document sources which need to be inspected, but a necessity for the correct application of normative principles, which should be pronounced, discussed and dealt as monolithic utterances instead of being sparse across several distinct codes.

The first step to accomplish the objective of legal simplification is to identify relations of pertinence between distinct laws, so that these can be unified and reproduced in new synthetic codes.

The research from which this work originated, which has been conducted in a collaboration framework between CNIPA (Centro Nazionale per l'Informatica nella Pubblica Amministrazione) and the ART group of the University of Rome "Tor Vergata", is addressing two complementary goals: to facilitate search and retrieval of information over large corpora of legal documents, and to support the process of

legislative simplification (which is being applied now in Italy) by detecting semantic associations between different legal sources.

In this paper we present STIA, a semantic annotation tool developed inside the above collaboration framework, with the objective of supporting legal professionals in exploring and cross-annotating the above semantic associations between heterogeneous legal sources.

STIA allows domain experts (lawyers, administrative staff, researchers,…) to inspect – through an ordinary web browser – laws, sections and paragraphs from two different electronic sources (web sites, digital repositories etc…), and to compare their content aiming to annotate relations of pertinence between them. STIA has been developed as an extension of a wider framework for knowledge acquisition and management based on the Firefox Web Browser [1], called Semantic Turkey [2][3].

1. State of the art

The information management in legal domain is a very important and well assessed research field. Several European countries have launched institutional projects to manage legislative corpus of Europe and in particular for the definition of identification standards for legal sources:

- In England legal simplification has been the aim of project LAMS (Legal and Advice Sector Metadata)[8]).
- For Italy, we can cite the project "Norme in Rete" [9] , coordinated by the Ministry of Justice.
- In the Netherlands the project Metalex, at the University of Amsterdam[10].

Furthermore, there are some international project targeted to manage many international corpora of legal acts:

- In Germany: the Lexml project [11],
- In USA: the Legalxml [12].

Finally, in Italy, the project "*JurWordNet*" [13] aims at realizing a semantic resource for Italian legal documents, containing a series of specific terms for the jurisprudence domain.

All these efforts are not yet enough: a recent user survey realized in the framework of the European funded Judicial Management by Digital Libraries Semantics (JUMAS) project [14] offers a quite clear view of the need for effective retrieval of textual and multimedia documentation in the criminal courts. More than 40% of the users evaluated the effectiveness of current tools for retrieving documentation or searching of relevant information as "poor" or critical.

In jurisprudence domain document collections tend to assume huge proportions (e.g. the Italian system of laws is composed by more than one hundred thousand of different acts). The biggest problem in this case is to retrieve useful information in such enormous collections in relatively short time.

Information Retrieval is typically used to retrieve relevant information, from a document collection. The matching between queries and documents is mostly term-based, i.e. the words within documents are used to describe the documents and to

determine their relevance for a given query. Moreover an expert in the legal domain needs more information than the simple correspondence between words.

To introduce more information about the meaning of a document, semantic annotations can be added, containing additional information about the text or part of it, that are important to improve retrieval processes. [15][16] [17].

In recent years, collaborative tagging systems have become very popular among users as a means for organizing their resources. These systems use semantic annotations taken by users to improve retrieval by using the information held into them. [18]

Due to the complexity and the vast scope of these issues a new discipline has born: the *cyberlaw*, dealing with, among others, management of metadata legal documents [19].

2. Knowledge Model

The knowledge model of the framework extended by STIA offers two concept layers, consisting in: the application layer, containing ontologies from Semantic Turkey and its extensions, which are necessary to drive the application, and the user layer, containing specific domain ontologies and allowing the user to add instance data.

In STIA, the application layer is constituted of:

- The *annotation ontology*: used by Semantic Turkey, provides the concepts for describing semantic annotations taken from text
- The *STIA application ontology*: STIA adopts a specific ontology (STIA Ontology from now on) for handling concepts from jurisprudence and those needed for the annotation (e.g. laws, constraint relationship between different part of law, and some relevant relationship between part of laws), and provides dedicated graphical interface for managing them. Note that, with respect to ST, no explicit representation of ontologies is given(i.e., in terms of classes, properties, instances), and the ontology model is only adopted to orchestrate the necessary information. Furthermore it is possible for users, by using the ontology editor features of Semantic Turkey, to add resources representing new similarity relationships and to delete existing ones, and these changes will be dynamically accounted into STIA.

The relationships of STIA Ontology can be divided in two main classes, according to the kind of investigation which is carried on by the user: relationships between different parts of the same law and relationships between parts of different laws.

The first class of relationships is part of the STIA application ontology, and is hidden from the user. Instance data for this ontology is already available (i.e. it provides containment semantics between laws, sections and paragraphs). This information is however implicitly available when the user browses through the laws.

The latter kind of relationships was defined by CNIPA domain experts and is thus part of the *domain ontology* of STIA (i.e. they are explicitly shown to the user, which can use them to annotate the text). The experts have defined by first the following relationships:

- *"Attuazione tramite uso strumentale dell'ICT per il raggiungimento di scopi più generali"* (in English: "Implementation through instrumental use of ICT for the achievement of objectives more general")
- *"Attuazione tramite uso dell'ICT per l'attuazione di principi astratti"* (in English: "Implementation through use of ICT for the implementation of abstract principle")
- *"Specificazione delle modalità d'uso dell'ICT per l'attuazione di principi astratti"* ("Specification of mode of use of ICT for the implementation of abstract principles")
- *"Specializzazione strumentale – uso dell'ICT per il raggiungimento di scopi più generali"* ("Specialization instrumental use of ICT for the achievement of objectives more general")
- *"Relazione esplicita inversa Attuazione tramite uso strumentale dell'ICT"* ("Explicit Inverse Relation, implementation through instrumental use of ICT")

These relationships will support users to annotate important issues concerning interactions between the various laws. STIA has been developed to allow users to annotate and manage these relationships, however, the set of relationships may be changed to reflect different needs and exigencies.

Semantic Annotations taken through STIA can be used both as a resource on their own (which can be used inside semantic search engines to easy retrieve related laws) as well as a precious feed for tools, based on machine learning techniques, aiming at automatically extracting more of the above relationships. Inside our collaboration with CNIPA, this task is being carried on through the "Naviga Norme"[7] tool

3. System Design

STIA is part of a framework that is composed of the already cited Semantic Turkey, and of two of its extensions: Range Annotator and XPointerlib.

Semantic Turkey (ST from now on) is a Semantic Web platform for Knowledge Management and Acquisition, realized by the ART Research Group[1] at the University of Rome, Tor Vergata. ST provides Ontology Development capabilities and facilitates the population of ontologies with new data by acquiring it from the Web. Through ST, users can literally select textual information from Web Pages, drag & drop it over ontology definitions to semi-automatically generate ontological data. ST offers a versatile extension mechanism combining OSGi standard [17] and Mozilla extension support thus allowing for the creation of completely new application residing on ST and on the hosting web browser.

Range Annotator [18] is an extension of ST replacing the standard annotation mechanism with one producing "RangeAnnotations". ST annotation mechanism produces in fact "semantic bookmarks", i.e. it keeps track of annotated pages, of their association to ontology resources and of the textual occurrences of these resources in the page. "Range Annotation" instead includes *range information* (that is: a *location* in the text defined by two points, a start point and an end point). This range

[1] http://art.uniroma2.it

information can be implemented according to different formats and interpreted accordingly by a dedicated annotation extension.

Taking pointwise Semantic Annotations is an important feature for STIA, which requires domain experts to be able to locate precise references inside law descriptions; currently, the pointer representation adopted in RangeAnnotator (and thus in STIA) is the W3C standard xpointer, which stores DOM information about the HTML Node(s) where the text has been annotated and its offset (in characters) from the start of the Node. In such a way we represent exactly the sentence extensions involved in the relations.

The RangeAnnotator extension implements the RangeAnnotation concept by adopting Xpointers standard to represent textual areas in the web page. The Xpointerlib [5] of Firefox (originally developed for another Semantic Web project: Annotea, to develop a Semantic Annotation platform for Firefox called Annozilla [20]) is used to identify xpointers in the text.

STIA further changed the standard annotation mechanism by allowing double pointers linking entries from the two inspected laws and a reference to the kind of relation existing between them, thus establishing a triple of constituents for the annotation.

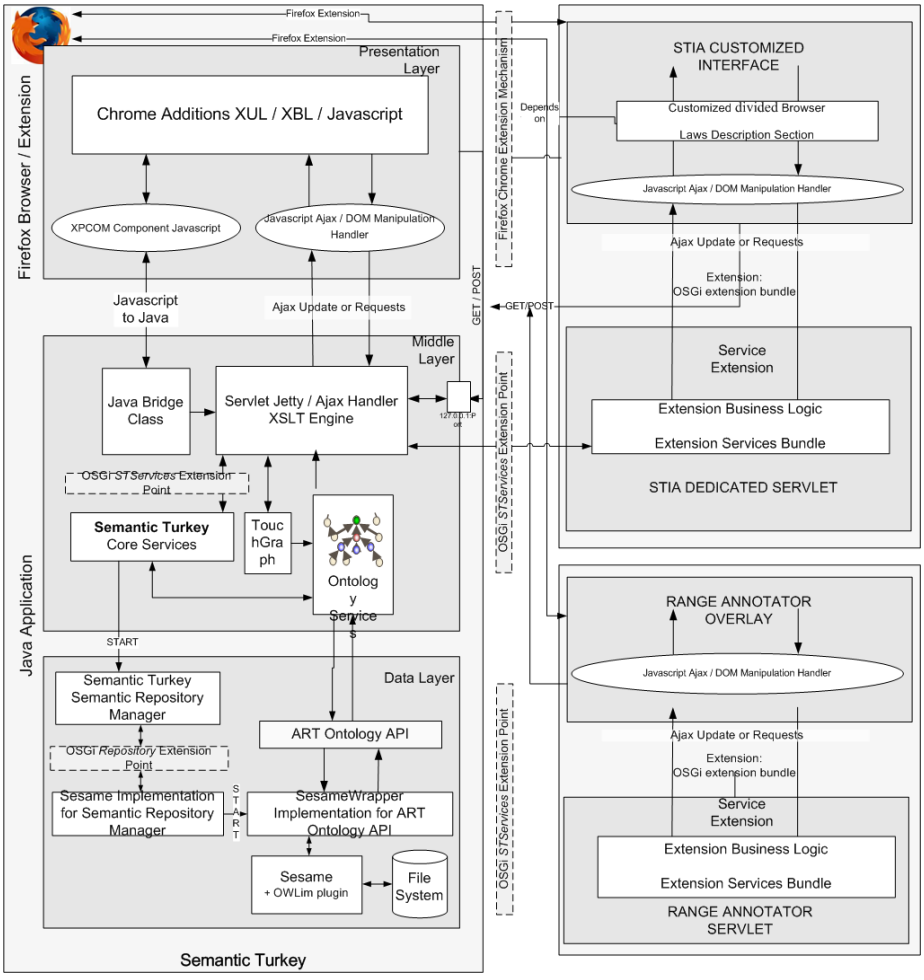

Figure 1. STIA Architecture: Semantic Turkey (on the left) and STIA and Range Annotator Extensions

Completely focused on this particular environment, the UI of the tool is totally original and does not extend the one offered by Semantic Turkey (which resembles more a standard ontology development tool such as Protégé [21]): it hides all the ontology editing capabilities of Semantic Turkey, though they are still available for "advanced" users.

The design of interface component was developed using feedback of CNIPA domain experts (that are the end users of STIA). The objective of this incremental design is to improve the usability of the interface and to make the task of annotation easier and faster.

4. Architecture

STIA is deployed as an xpi (cross-platform installers) package which, once installed inside Firefox[1], is handled by Semantic Turkey extension discovery system, which extracts OSGi bundles and installs them in the main application.

Figure. 1 shows the Architecture of Semantic Turkey (on the left) and its interfacing with STIA (right).

The architecture of *STIA* is composed of two main components: the user interface (UI) and dedicated services.

The UI sits on top of Semantic Turkey, by exploiting Mozilla overlaying mechanism (through which it is possible for new extension to "overwrite" the content of those on which they depend) and provides a completely new presentation layer.

STIA services provide new functionalities to: store annotated relationship, retrieve relationship related on a law, remove previously stored annotation and manage information about structure of law to populate interface.

STIA use the services functionalities of Range Annotator to take and store the semantic annotations.

5. User Interaction

STIA UI (shown in **Figure 2**) is divided into two main sections: the information panel and the browser panel. The STIA UI shows the two laws on which the user can browse laws and check their details.

The information panel (upper part of Figure 2**)** contains all the details of considered laws; the list of law sections is automatically filled with those from the selected law and the same is done for the paragraph list when the section is selected. Further, the user can choose the relationship that he wants to establish between the two parts by selecting one of the proposed relation types in the menubox labeled as "*Tipologia di Relazione"* (relation type). Finally, there are two text boxes containing the text that qualifies the relation: these are automatically filled when the user selects a section of text and presses the *"Testo annotato"* (annotated text) button into STIA Browser.

The browser of STIA (bottom section of Figure 2) is divided into two panels showing the two laws which have been previously loaded through the upper panel. It is also possible to navigate backward or forward in the visited pages as in any

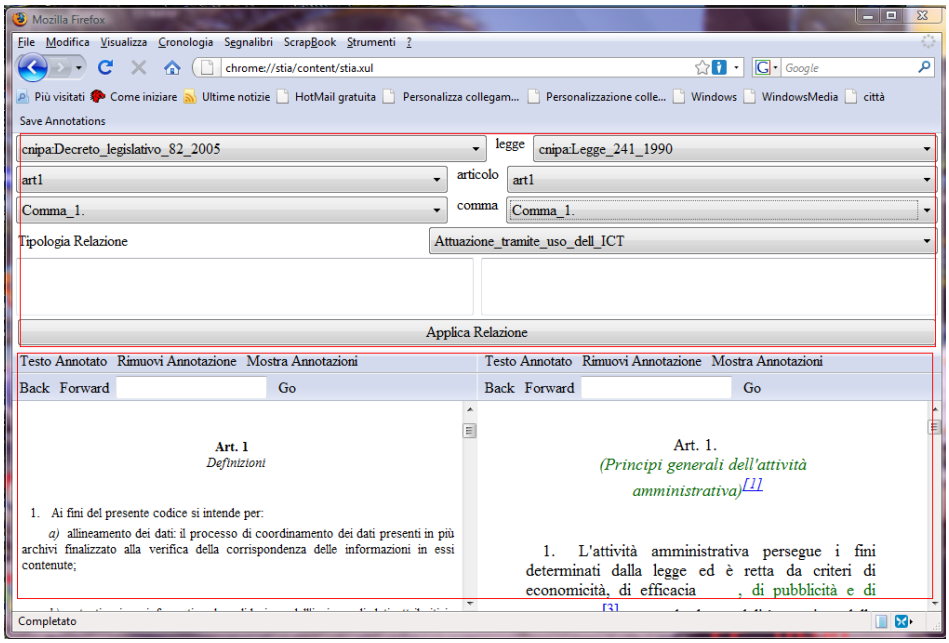

Figure. 2 STIA Main Interface.

traditional browser (so that the user is already well acquainted with the traditional web navigation); in this case, the upper panel is updated accordingly.

As shown in Figure. 3, in STIA it is possible to highlight the semantic annotation that were previously taken in a document, by depressing the "*Mostra Annotazioni*"(show annotations) button. This is very useful to simplify the work of annotators.

Supervisors can check annotated relationship and delete them if needed just by depressing the "*Rimuovi Annotazione*" (remove annotation) button.

Thus, when a user wants to take an annotation, he has to perform the following operations:

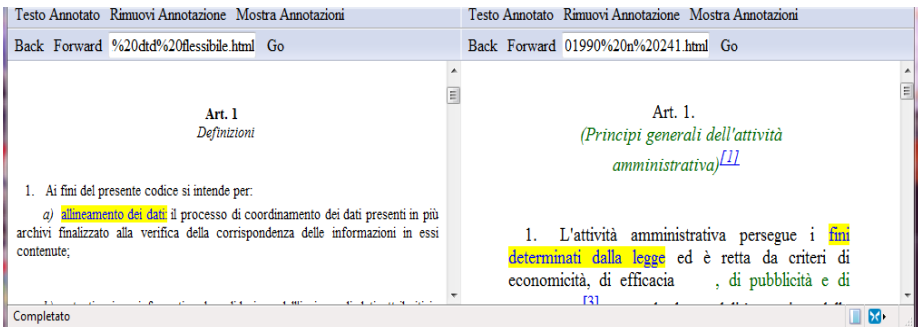

Figure. 3 Highlighted Semantic Annotation in STIA Browser.

1. Opens the information panel
2. chooses the laws he wishes to browser
3. finds a relationship between the browsed law
4. defines the type of relationship he wishes to assert (such as "*Attuazione tramite l'uso dell'ICT*" in Fig. 2)
5. on the browser panel he navigates into the previously selected laws and selects the part of the whole document that is relevant for the considered relationship
6. finally the user confirms the relationship by depressing the "Applica Relazione" (*annotate relationship*) button.

6. Future Works

The development of STIA is a first fundamental step of a complex framework that will allow legal domain experts to manage legal corpora in a fast way and to annotate several information over them.

This framework could be used in several application contexts to simplify the writing of new laws, in this case the lawmaker needs many information about previous laws that are in relation with the new law that he is willing to introduce.

The semantic annotations taken with STIA can be used in two ways: as a resource on their own, feeding semantic search engines supporting lawmakers and legal experts in general which may need to easily browse through related laws, as well as a collection of annotations which may be used to train machine learning tools to extract further relationships of the same type on new texts.

Regarding the second use of the tool, STIA was developed in parallel with the "Naviga Norme" tool (CNIPA). "Naviga Norme" tool is a platform allowing specialists of the legal domain to retrieve general relationships between paragraphs of normative texts. In particular "Naviga Norme" can take one paragraph as input and use it to make a query. "Naviga Norme" then returns a list of paragraphs that are in relationship with the initial one, sorted by score. We have started to use STIA to improve the retrieval of the relationship between laws or part of them introducing qualified relationship and using the semantic annotation taken using STIA to modify the ranking of "Naviga Norme" tool.

The semantic annotations taken with STIA could be used by "Naviga Norme" to expand the original query automatically and to improve the IR process.

Moreover they could be activate a machine learning process to train the system and identified new instances of the relationship defined in STIA Ontology which are not yet been annotate from the users.

We can assume a scenario in which the user will load a document in STIA and the system will answer by showing the metadata already included, by him or by another user, using the instrument of semantic annotation offer by STIA if these are available.

Furthermore, the system will answer by showing new proposals of semantic annotations that the user could confirm or reject for inclusion in knowledge base.

7. Conclusion

In this paper we described STIA: a semantic annotation tool customized for being used in the jurisprudence domain.

The graphical user interface of STIA has been incrementally designed and implemented using the feedback collected by the domain experts from CNIPA.

The goal of this incremental development process was to improve the usability of the interface and to make the task of annotation easier and faster.

At the end of the implementation phase, all the experts involved in the research considered the UI "very satisfying" since it well fits the way they think, organize the information and develop their works.

The Knowledge Model of STIA is designed especially to represent the legal domain and in particular the Italian laws system. Moreover STIA allows to change the STIA Ontology and increase it by using the features expressly developed into the system.

STIA simplifies and speeds up the usage of large legal document collections mainly enabling both the annotation of explicit semantic relations between fragments of normative texts and, consequently, by reusing collected annotations to browse through semantically interconnected laws.

References

1. Firefox home page. Available at: http://www.mozilla.com/en-US/firefox/
2. Griesi, Donato, Pazienza, MariaTeresa, Stellato, Armando: Gobbleing over the Web with Semantic Turkey. In : Semantic Web Applications and Perspectives, 3rd Italian Semantic Web Workshop (SWAP2006) (2006)
3. Pazienza, MariaTeresa, Scarpato, Noemi, Stellato, Armando, Turbati, Andrea: Din din! The (Semantic) Turkey is served! In : SWAP 2008, Roma (2008)
4. http://www.lcd.gov.uk/consult/meta/metafr.htm#part6.
5. http://www.normeinrete.it/. In: http://www.normeinrete.it/.
6. http://www.metalex.nl/pages/welcome.html. In: (http://www.metalex.nl.
7. http://www.lexml.de/.
8. http://www.legalxml.org.
9. http://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/JurWordNetEng.htm.
10. Judicial Management by Digital Libraries Semantics (JUMAS) Project homepage. In: http : //www.jumasproject.eu/.
11. Lioma, Christina, Moens, Marie-Francine, Azzopardi, Leif: Collaborative Annotation for Pseudo Relevance. In : ECIR'08 Workshop
12. Lesmo, Leonardo, Mazzei, Alessandro, Radicioni, Daniele: Extracting Semantic Annotations from Legal Texts. In : HT '09: Proceedings of the Twentieth ACM Conference on Hypertext and HypermediaNew York, NY, USA: ACM , July (2009) (2009)
13. Bartolini, Roberto, Lenci, Alessandro, Montemagni, Simonetta, Pirrell, Vito, Soria, iClaudia: Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study. In : Second International Workshop on Regulatory Ontologies (2004)
14. Yeung, Ching-man, Gibbins, Nicholas, Shadbolt, Nigel: Web Search Disambiguation by Collaborative Tagging. In : Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'08) on 30th European Conference on Information Retrieval (ECIR 2008) (2008)

15. Fioriglio: Temi d' informatica giuridica. In: http://www.computersworld.eu/download/fioriglio-temi_di_informatica_giuridica.pdf. (Accessed 2004)
16. CNIPA: http://www.cnipa.gov.it/. In: http://www.cnipa.gov.it/.
17. OSGi: OSGi Bundle Repository Specification. In: OSGi RFC0112. (Accessed 2005) Available at: http://www2.osgi.org/Download/File?url=/download/rfc-0112_BundleRepository.pdf
18. Art Group Tor Vergata: Range Annotator: Semantic Annotation on Semantic Turkey. In: http://semanticturkey.uniroma2.it/extensions/rangeannotator/.
19. http://xpointerlib.mozdev.org/. In: XPointerLib.
20. Wilson, Matthew: Annozilla (Annotea on Mozilla). In: http://annozilla.mozdev.org/.
21. Gennari, John, Musen, Mark, Fergerson, Ray, Grosso, W, Crubézy, Monica, Eriksson, H., Noy, Natalya, Tu, Samson: The evolution of Protégé-2000: An environment for knowledge-based systems development. International Journal of Human-Computer Studies 58(1), 89–123 (2003) Protege.
22. Art Group Tor Vergata: www.art.uniroma2.it. In: Art Group.