UNIVERSITY
OF TRENTO

**DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE**

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

# USING SEMANTIC ANNOTATION FOR MINING PRIVACY AND SECURITY REQUIREMENTS FROM EUROPEAN UNION DIRECTIVES

Paolo Guarda, Nadzeya Kiyavitskaya and Nicola Zannone

February 2008

Technical Report # DISI-08-011

# Using Semantic Annotation for Mining Privacy and Security Requirements from European Union Directives

Paolo Guarda, Nadzeya Kiyavitskaya, and Nicola Zannone

Dept. of Legal Sciences and
Dept. of Information Science and Engineering, University of Trento, Italy, and
Dept. of Computer Science, University of Toronto, Canada

**Abstract.** The increasing complexity of software systems and growing demand for regulations compliance require effective methods and tools to support requirements analysts activities. In order to facilitate alignment of software system requirements and regulations, systematic methods and tools automating regulations analysis must be developed. This work explores applicability of the semantic annotation tool Cerno to mining of rights and obligations from European privacy directives.

## 1 Introduction

Security, privacy and governance are increasingly the focus of government regulations in Europe and elsewhere. Among these, the special concern is drawn to the regulations on privacy given the growing importance of appropriate processing of private and sensitive information on the Web. To this end, the European Union (EU) has issued several directives on privacy that contain general guidelines for processing personal data. These directives must be implemented by each member state of the EU. This situation has created the regulation compliance problem, whereby companies and developers are required to ensure that their software systems comply with relevant regulations, either through design or reengineering.

Accordingly, in this work we aim to further bridge the gap between information technologies and the domain of legal documents, thus providing a better support for software engineers in devising high quality software systems that would be compliant with both national and community laws.

Acquiring requirements from regulation documents is a challenging task for requirements engineers [8]. Invariably, these regulations are specified in textual format. The difficulty lays in the nature of these texts. Regulations are written in natural language, use legal terms and are laden with ambiguities - a pervasive phenomenon with natural languages [7].

The process we envision for extracting requirements from regulations consists of three steps:

1. regulatory text is annotated to identify text fragments describing actors, rights, obligations, etc.;

2. a semantic model is constructed from these annotations; and
3. the semantic model is transformed into a set of functional and nonfunctional requirements.

The first two steps are currently supported by Breaux and Antóns systematic, manual process for deriving semantic models from policies and regulations called Semantic Parameterization [8]. In this process, rights and obligations from regulation texts are first restated into restricted natural language statements (RNLS) and then mapped into formal semantic models. Extracted semantic models can be queried and analyzed for ambiguities and conflicts. Each RNLS should describe a single activity with external references to other activities. The statement has exactly one primary actor, action and at least one object.

Previously, we proposed to provide a tool support for this methodology [15] using as the baseline technology for analysis of legal documents the semantic annotation tool Cerno [13]. Cerno accepts as input a grammar and a document, generates a parse tree for the input document, and applies transformation rules to generate output in a target format. The approach discriminates between domain-dependent and independent components of the annotation process and thus allows for easy adaptation to different application domains and tasks.

In the present work, we extend and generalize the Cerno framework for the analysis of a wider range of legal documents. More specifically, we focus on the European privacy directives. The contributions of this work includes a database backend to the Cerno semantic annotation framework. To realize this feature we used the method for querying XML documents of Atre [5].

This paper is structured as follows. The baseline of the present work in sketched in Section 2. It introduces Cerno-based process for semantic annotation of legal documents. Section 3 discusses the difficulties of mining software requirements from European directives. Section 4 describes how the baseline technologies were extended to cater for the specifics of the European directives. Section 5 presents the setup and evaluation of the case study and summarizes the lessons learned. Section 6 recalls the related work. Finally, the conclusions are drawn in Section 7.

## 2 Cerno-based Process for Regulation Analysis

The tool-supported process for regulation analysis that we previously developed is based on the methodology for extracting stakeholder requirements from regulations by Breaux et al. [8], see Fig. 1.

According to the methodology, the process for extracting requirements from regulations consists of three steps:

- regulatory text is annotated to identify text fragments describing actors, rights, obligations, etc.;
- semi-formal rights, obligations and constraints are formally modeled in first-order predicate logic using a process called Semantic Parameterization that provides increased precision; after that, the semantic model can be analyzed for inconsistencies and corrected by an expert;

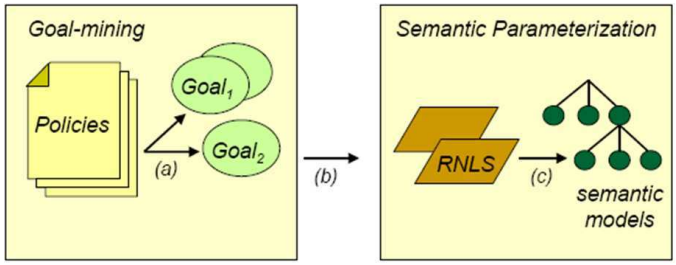


**Fig. 1.** Manual methodology for extracting requirements from regulations

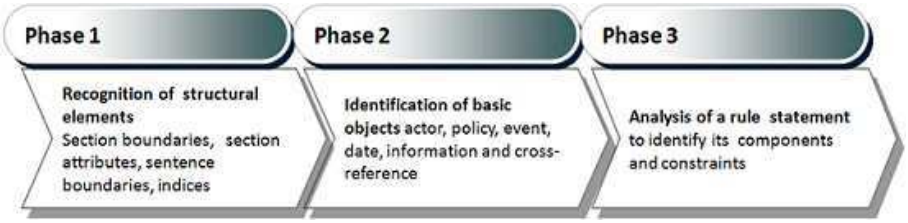


**Fig. 2.** Cerno-based regulation analysis

– the semantic model is transformed into a set of functional and nonfunctional requirements.

The tool we previously developed to support this methodology on the example of the U.S. privacy rule HIPAA [19] recognizes document structure in terms of section and subsection boundaries, titles and annotated paragraph indices, identifies instances of the concepts actor, policy, event, information and date and annotates document fragments describing *rights, anti-rights, obligations, anti-obligations,* and related *constraints* [15]. To generate these annotations, the tool used a list of normative phrases for the objects of concern that was obtained by manual analysis of the HIPAA document [8].

In a nutshell, the regulation analysis process consists of three main phases [15], as shown in Fig. 2:

- Recognition of structural elements of the document: section boundaries, section attributes which are number and title, sentence boundaries;
- Identification of basic objects: actor, policy, event, date, information and cross-reference;
- Deconstruction of a rule statement to identify its components and constraints.

This process is based on Cerno [13], a lightweight semantic annotation framework that exploits fast and scalable techniques from the software reverse engineering area, more specifically, "design recovery" process [10]. To annotate input

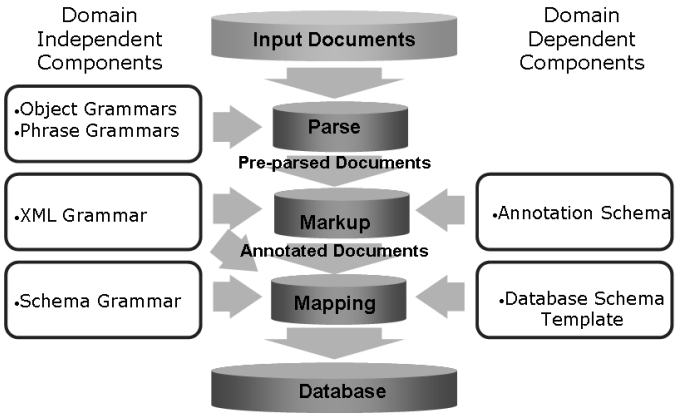


**Fig. 3.** The workflow in the Cerno's semantic annotation framework

documents, Cerno uses context-free grammars, generates a parse tree, and applies transformation rules to generate output in a target format [12].

The generic architecture of Cerno (Fig. 3) consists of a number of consequent transformations:

1. *Parse.* The tool parses an input document breaking it down into its constituents according to a predefined document grammar that is domain independent. The produced parse tree consists of structures such as document, paragraph, phrase, and word. Any of these structures can be chosen as an annotation unit, depending on the purpose of annotations. At the same time, complex word-equivalent objects, such as phone numbers, e-mail and web addresses, and similar structures, are properly recognized using structural patterns of object grammars. Grammars are described in a BNF-like form.
2. *Markup.* This stage uses a domain-dependent annotation schema to infer annotations. An *annotation schema* is a specification of the kinds of annotations to be generated. It is composed of a set of semantic tags along with their syntactic indicators: positive, which point to the presence of a concept instance, and optionally negative, which on the contrary exclude its presence, i.e., they are counter-indicators. These domain-specific indicators can be derived manually, i.e., proposed by the domain experts, or semi-automatically, for instance, mined from a rich conceptual model representing the domain knowledge if such a model is available. The processing exploits the structural pattern matching and source transformation capabilities. *Syntactic indicators* can be contain literal words and phrases, or names of parsed entities.
3. *Mapping.* This stage is optional. It is executed in the case analysts have to store extracted information in an external database. In this stage, annotated fragments are selected from all annotations according to a predefined database schema template and, then, are inserted into the database. The

schema is domain dependent and represents a sort of a target template to accommodate annotations relevant to a specific task.

## 3 European Privacy Directives

In the present work we focus on the analysis of two European Union privacy directives:

- *Directive 95/46/EC* of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1];
- *Directive 2002/58/EC* of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [2].

Thus, the first directive describes the rules for protection of privacy of an individual and the second one complements the earlier directive by specifying the rules for privacy on electronic communications for individuals, legal persons, and organizations.

These data protection regulations set the legal principles and requirements that must be met by organizations when processing personal data. The privacy can be reached only if the system is built up so as to protect it. In particular, these regulations include:

- Definition of general principles with regard to the processing modalities,
- Acknowledgment of specific rights to every data subjects,
- Specific regulation of the so called "sensitive data".

One must take into account that, unlike in regulations, obligations and rights contained in the European directives do not specify the exact way of implementing them. The directives are intended to define the desired result leaving to a Member State some freedom in choosing a particular way, which better fits in the national legal tradition, to implement this result.

Consider, for instance, this fragment of the Directive 95/46/EC: "`Member States may provide that data relating to administrative sanctions or judgements in civil cases shall also be processed under the control of official authority.`" The above statement does not specify a right from a legal point of view, but actually represents a sort of possibility left to the Member State in order to harmonize and implement the requirement in the specific national legal system. However, for the sake of simplicity of our application, we equalize such statements with rights.

[1] http://europa.eu/scadplus/leg/en/lvb/l14012.htm

[2] http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdocnumdoc=32002L0058 model=guichettlg=en

## 4 Research Methodology

Similarly to the analysis of HIPAA [15], we incorporated a set of the "concepts of interest" as the annotation schema for Cerno:

- *actor*, *policy*, *resource*, and *action*: at words level,
- *obligation* and *right*: at sentence level,
- *condition*, *refinement*, *exception*: at phrase level.

where:

- An *actor* is a natural or legal person or organization that is involved in the action.
- A *policy* can be the name of the law, standard, act or other regulation document which establishes rights and obligations.
- A *resource* is a physical or informational entity that is of special value.
- *Actions* are verb phrases involved in rule statements and describe *what* a stakeholder must of is allowed to do.
- A *right* is an action that a stakeholder is conditionally permitted to perform, e.g., "`An entity may request an extension of the preparation period`".
- An *obligation* is an action that a stakeholder is conditionally required to perform, e.g., "`The Responsible Entity shall maintain a list of designated personnel who are responsible for authorizing logical or physical access to protected information`".
- A *condition* is the part of a rule statement that describes a situation of the rule applicability, for instance, "`where external interactive access into the Electronic Security Perimeter has been enabled`".
- *Refinement constraints* elaborate the domain by listing its specializations, as for instance the phrase "who are responsible for authorizing logical or physical access to protected information" from the above example.
- *Exceptions* remove elements from consideration in a domain, e.g., "`unless subscribers indicate otherwise`" or "`except when legally authorised to do so`".

Note, that in the present work we do not annotate anti-rights and anti-obligations only because explicit instances of these concepts were not identified neither manually nor automatically in the text of directives.

To recognize instances of the actor, policy and resource concepts, we can exploit the regularity of the document, meaning that as in most regulations and policies, the directives use standard terms and limits the use of synonyms to the definitions of those terms. Normally, any legal document contains an introduction section where every used term is strictly defined and a reference synonym is assigned. Fig. 4 shows the indicator lists for basic concepts. Symbol "|" is used to list alternative choices for the parser.

Next, in order to identify action verbs, we've been exploiting the results provided by a Part of Speech Tagger (POS) [18] by drawing a list of all verbs in present tense from the text of both directives.

**Actor**: the council of the european union, user of a publicly available electronic communications service, provider of a publicly available electronic communications service, legal person(s), data (subject(s) | controller(s)), supplier(s), service provider(s), provider(s), member state(s), third (party| parties | country | countries), subscriber(s), user(s), controller(s), processor(s), recipient(s), commission, european parliament, council, operator(s), person(s), people, customer(s), working party, working parties, (supervisory | official | public) authority, (supervisory | official | public) authorities;
**Resource**: (private | user | sensitive | location | traffic | personal) data;
**Policy**: community law, european convention for the protection of human rights and fundamental freedoms;

**Fig. 4.** The list of indicators for basic concepts derived from the definition sections of two directives

**Table 1.** Normative phrases for the concept of Right in past and present applications

| No | Right | Status |
|---|---|---|
| 1 | <actor> may | |
| 2 | <actor> can | |
| 3 | <actor> could | |
| 4 | <policy> permits | |
| 5 | <actor> has a right to | |
| 6 | <actor> should be able to | |
| 7 | <actor> shall be given the possibility | New |
| 8 | <actor> must continue to have the possibility | New |
| 9 | <actor> shall have the right to | New |
| 10 | <actor> must have this possibility | New |

Integrating all the considered heuristics into the Cerno's domain dependent components, we complete the first step of the regulation analysis process [13], i.e., annotation of basic entities at the words level. On the basis of the processing the more complicated rules, combining contextual keywords and names of recognized basic entities, can then be applied to identify complex concepts.

Consequently, for identification of such concepts as rights, obligations and various types of constraints, we were able to partially reuse the result of Breaux and Antón, where a list of normative phrases for these concepts was derived by manually analyzing the HIPAA document [8]. In addition to the phrases identified in their work, we extended Cerno by new heuristic rules that were found useful, see for instance the extended list of normative phrases for the concept of right in Table 1.

By applying the rules based on the defined normative phrases, we fulfill the second step of the regulation analysis process, i.e., identification of complex entities at the sentence and phrase level. See a fragment of the annotated document of Directive 95/46/EC in Fig. 5.

<policy>**Article 18**</policy>
Obligation to notify the <actor>supervisory authority</actor>
1.<Obligation> <stakeholder><actor>Member
States</actor></stakeholder> shall <action>provide</action>
that the <actor>controller</actor> or his representative,
<Condition>if any</Condition>, must <action>notify</action>
the <actor>supervisory authority</actor> referred to in
<policy>Article 28 </policy>before carrying out any wholly or
partly automatic processing operation or set of such operations
intended to <action>serve</action> a single purpose or several
related purposes</Obligation>.
2.<Right> <stakeholder><actor>Member States</actor></stakeholder>
may <action>provide</action> for the simplification of or
exemption from notification only in the following cases and under
the following conditions:
- where, for categories of processing operations which are
unlikely, taking account of the data to be processed, to
affect adversely the rights and freedoms of <actor>data
subjects</actor>, they specify the purposes of the
processing, the data or categories of data undergoing
processing, the category or categories of <actor>data
subject</actor>, the <actor>recipients</actor> or categories
of <actor>recipient</actor> <Refinement>to whom the data
are to be disclosed and the length of time the data are to be
stored</Refinement>, and/or
- <Condition>where the <actor>controller</actor></Condition>,
in compliance with the national law which governs him, appoints
a <resource>personal data</resource> protection official,
responsible in particular:
- for ensuring in an independent manner the internal application
of the national provisions taken pursuant to <policy>this
Directive</policy>
- for keeping the register of processing operations carried out
by the <actor>controller</actor>, containing the items of
information referred to in <policy>Article 21 (2)</policy>,
thereby ensuring that the rights and freedoms of the <actor>data
subjects</actor> are unlikely to be adversely affected by the
processing operations</Right>.

**Fig. 5.** An annotated extract from Directive 95/46/EC

This example contains two rule statements: an obligation and a right. In both statements the tool identified all actors involved, names of other regulative documents, i.e. instances of the policy concept, and one resource, "`personal data`". One of the actors was marked up as stakeholder in each of the statements. Inferred annotations include instance of refinement and condition. In the obligation we may notice that the word "`serve`" has been incorrectly annotated as

an action verb. In the given right, the condition "`where <...> they specify the purposes of the processing ...`" was missed by the tool.

To wrap the annotated results in such a way that they can be further queried and analyzed for completeness, we utilized the work of Atre [5], who uses source transformation techniques to transform an XML document into SQL statements given a set of desired tags. When the Atre's approach is applied, the generated SQL statements can be then executed on a database server to create a relational view over the XML document. This view can then be queried using SQL to get information in the XML document. The approach requires a user to define a set of XML element for the resulting database tables in this way avoiding the storage of instances irrelevant for the purposes of a specific application. This reduces search scope and makes querying more focused. The approach is independent of the backend database system.

As a result of applying the Atre's approach, we generated a MS Access database containing the annotated instances of the concepts of interest. Populating the database with the annotated text fragments completes the final step of the Cerno framework. Table 2 shows a fragment of the retrieved data for a sample query on the extracted data, i.e. "show all rights that are restricted by some conditions".

## 5 Experimental Case Study

To verify generality of the Cerno-based approach, we applied this process on the text of two European directives: Directive 95/46/EC, containing a total of 12682 words, and Directive 2002/58/EC, containing a total of 8552 words. As a result, the full text of these documents was annotated with a total of 1295 and 882 tags for the former and the later directives respectively. The processing times were about 1.5 for Directive 95/46/EC and 1.1 seconds for the one of 2002 on Intel Pentium 4, 2.60GHz, Ram 512 MB, running Windows XP.

### 5.1 Evaluation Method

We evaluated the performance by comparing automated results to a *Gold model*, i.e. the annotation drawn manually by the experts, and calculating recall and precision quality measures [6]. Let $TP$ be the number of true positives, i.e. relevant items retrieved, $FP$ – the number of false positives, i.e. irrelevant items retrieved, $FN$ – the number of false negatives, i.e. relevant items missed, and $TP + FP$ – the total number of retrieved items. Then, the quality measures are defined as follows:

- *Recall* shows how well the tool performs in finding relevant items and is calculated by dividing the number of relevant items found by the number of all relevant items in the collection: $Recall = TP / (TP + FN)$;
- *Precision* shows how well the tool performs in not returning irrelevant items and is produced by dividing the number of relevant items found by the number of all items found: $Precision = TP / (TP + FP)$.

**Table 2.** Query answering based on the information annotated by Cerno

| **tagId** | **Right.tag_content** | **Condition.tag_content** |
|---|---|---|
| *Right1* | The service provider may process traffic data relating to subscribers and users where necessary in individual cases in order to detect technical failure or errors in the transmission of communications | where necessary in individual cases in order to detect technical failure or errors in the transmission of communications |
| *Right10* | Where consent of the users or subscribers has been obtained for the processing of location data other than traffic data, the user or subscriber must continue to have the possibility, using a simple means and free of charge, of temporarily refusing <...> | Where consent of the users or subscribers has been obtained for the processing of location data other than traffic data |
| *Right11* | Member States may require that for any purpose of a public directory other than the search of contact details of persons on the basis of their name and, where necessary, a minimum of other identifiers, additional consent be asked of the subscribers | where necessary |
| *Right13* | Member States may adopt legislative measures to restrict the scope of the rights and obligations provided for in Article 5, Article 6, Article 8 (1), (2), (3) and (4), and Article 9 of this Directive when such restriction constitutes a necessary, <...> | when such restriction constitutes a necessary |
| *Right3* | Member States may restrict the users' and subscribers' rights to privacy with regard to calling line identification where this is necessary to trace nuisance calls and with regard to calling line identification and location data where this is necessary to allow emergency services to carry out their tasks as effectively as possible | where this is necessary to trace nuisance calls and with regard to calling line identification and location data<br><br>where this is necessary to allow emergency services to carry out their tasks as effectively as possible |

**Table 3.** Evaluation summary

| | **Directive 2002** | | **Directive 1995** | |
|---|---|---|---|---|
| **Concept** | *Recall* | *Precision* | *Recall* | *Precision* |
| exception | 1.00 | 1.00 | 1.00 | 1.00 |
| condition | 0.97 | 1.00 | 0.97 | 1.00 |
| refinement | 0.80 | 1.00 | 0.86 | 1.00 |
| stakeholder | 0.91 | 1.00 | 0.96 | 0.99 |
| action | 0.96 | 0.94 | 0.91 | 0.99 |
| Right | 0.75 | 1.00 | 1.00 | 1.00 |
| Obligation | 0.87 | 1.00 | 0.95 | 1.00 |
| **Average** | **0.90** | **0.99** | **0.95** | **1.00** |

### 5.2 Evaluation Results

For assessing the quality of automated annotations, two human experts manually checked the annotated text of both directives correcting tool's errors. Then, Recall and Precision measures for each concept were calculated. The performance values are presented in Table 3.

Overall, the evaluation results demonstrate very high performance scores, especially with respect to the precision measure.

Relatively low recall values were shown for some concepts, e.g., refinement and right. This was caused by the lack of heuristic rules catching all possible variations of their instances. Indeed, we initially adopted the set of normative phrases that would guarantee high precision for the related concepts and can't be confused with their usages for other concepts. In this way, annotations generated by using these normative phrases provide a sound starting basis for a software engineer that need to mine requirements from a given regulative document. The human then needs only to revise unannotated text fragments and find several missing instances in them. Another reason for incomplete retrieval for the concepts at the level of sentence sub-clauses, such as refinement and condition, lays in the occasional use of such sentence constructions where one phrase is split by one or more other phrases, for instance, as the condition "`where, for categories of processing operations which are unlikely, taking account of the data to be processed, to affect adversely the rights and freedoms of data subjects, they specify the purposes of the processing ...`" in Fig. 5. The tool was not able to detect such instances, as the heuristic rules adapted so far cannot generalize over underlying linguistic parse trees of sentences.

As for the concept of action and stakeholder, imperfect recall for these entities is mainly motivated by two reasons: (a) several usages of passive tense constructions that were not catch by the normative phrases, and (b) conjuncted actions or stakeholders are scattered over itemized subparagraphs. However, this drawback of the tool can be effectively addressed by prompting the user revising the automated annotations to fill in missing actions for identified instances of rights and obligations. Thus, the human won't need to read the entire document, but only to look up for the basic entities for selected statements.

Imperfect precision for actions is caused by several verbs which the tool has erroneously annotated in subordinate clauses apart from the actions of the main sentence clause of a right or an obligation itself. These false positives can be quickly identified and removed by a user who revises basic entities contained in a given requirements.

### 5.3 Discussion of the Results

The results speaks clearly in favour of the effectiveness of applying the Cerno-based process for the analysis of regulative documents. By applying this tool-supported process we were able to obtain semi-formal rights and obligations that can be then checked for consistency and represented as a set of formal requirements. The generated (functional) requirements can also serve as the basis for testing of software systems and thus verifying their compliance to privacy regulations.

It is important to note, that due to the nature of the European privacy directives analyzed in the present work, the requirements mined from them can be then reused in other domains that affect privacy issues in all Member States of the EU. Thus, the results of our work can be further applied in software development projects for modeling privacy requirements.

## 6 Related Work

Given the need in facilitating the work of legal experts developing high quality legislations, standards and policies, a number of methodologies and tools have been developed. Still, there is an urgent need in establishing a dialog between software developers and legal experts, so that both parties can benefit from better understanding of their needs.

In [2] Antón proposed the Goal-Based Requirements Acquisition Methodology (GBRAM) to manually extract goals from natural language documents. The GBRAM has since been applied to financial and healthcare privacy policies [3].

To facilitate reasoning with regulations, Antoniu et al. [4] introduced the regulations analysis method based on defeasible logic rules [16]. For this purpose, the facts manually found in the regulation document should be represented as a set of defeasible theory.

In order to improve accessibility to laws, by offering the support of legal drafting, several efforts are being realized. One of the products of these efforts is NormaSystem [17], a user-friendly tool for creating and annotating legal documents. The tool is implemented in Visual Basic .NET for Office XP and runs on top of the Microsoft Word. The system supports manual annotation activity. Acceptable input formats are HTML, XML, RDF, and plain text. The tool then validates annotated documents according to the document structure (DTD and XML-schema validation), thus detecting inconsistencies in the semantic markup, for example, a missing publication date, duplication of the title or date, wrong

content type. The tool also includes a converter for transformation of documents into different standards. In a way similar to the user-friendly manner of annotation, the Norma-System provides the possibility to update a legislative document using the consolidation module. In turn, the Norma-Server serves not only as a repository for documents and metadata, but also manages versioning and provides some facilities for legal reasoning. The authors claim that reasoning module detects conditional modifications and uses *defeasible logic* to represent them, although this mechanism is not clearly explained in the publications.

MetaVex [20] is a regulation-drafting environment intended to be used by drafters and member of parliament. For this purpose, it provides editing facilities in a WYSIWYG interface similar to a conventional word processor. The system is developed within the Java Eclipse platform. The user starts creating the content in a word processor. In this stage, a set of templates structured according to the Dutch Guidelines for Legal Drafting can be used to facilitate the composing process. Elements that are frequently used in the domain of legal documents, such as citation, appendix, titles, and others, are factored out in a separate panel. Each of them can be instantiated by the user in an appropriate position. The user can then manually modify metadata attributes. The tool allows the user to add metadata both to a document's fragments and to the document as a whole. In addition, it provides the possibility for marking references to elements of (other) regulations and to individual entities, such as institutions or concepts defined by the regulation. Documents are saved in XML that complies with the MetaLex[3] format for legal sources. The MetaVex environment is strongly connected with the existing semantic web standards, such as RDF Schema and OWL. A document produced by the environment can be converted into RDF or OWL by means of XSLT transformations. The tool is under construction as some of the promised features are not fully implemented yet.

XMLegesEditor [1] is a legislative drafting environment developed to facilitate the adoption of Italian Legislative National XML Standards (NIR). The authors of the tool argue that existing WYSIWYG word-processors mainly focus on *style* markup rather than on structural and *semantic* markup. Therefore the original solution is proposed. In addition to providing a traditional word processor for creating the document content, the tool *a priori* guarantees generation of a valid XML document by constraining the user to perform only valid operations on the document. In order to support annotation using NIR elements, the tool provides a toolbar containing such elements. The system is Open Source and written in Java.

The three editing systems considered above share many features. Each of them presents an original editing environment that allows legal experts to create and modify textual content in an understandable, transparent way. This means that users do not need to have any programming skills or knowledge of XML. XMLeges is principally oriented to developing documents according to the NIR standard. The modules of the system that automate semantic annotation of a legal document have been specifically trained to classify provisions and

[3] http://www.metalex.eu/wiki

identify instances of the elements of according to this standard in texts written in the Italian language. On the other hand, Norma-System and MetaVex are not biased to a specific language. Though MetaVex provides a possibility to use the template based on the Dutch standards for legal writing to facilitate human's work, other templates can be incorporated. All three systems generate valid XML document in the end of processing that can be then converted to other formats. However, the first two tools enable annotation of only simple semantic entities, such as author, date, title, and other similar information. Whereas the XMLegesEditor enables generation of a wider range of semantic metadata by automatically classifying provisions by their types and identifying arguments required by a particular type according to concepts of the NIR standard.

To this end, the Cerno-based annotation process is complement to the regulation-drafting environments. Because Cerno allows constructing a generic process for analysis of documents, this feature makes the tool applicable to different types of regulatory texts, semantic models and languages. However, Cerno lacks a back-end to the existing standards and, therefore, such environments, as for instance MetaVex and Norma-System, can serve as a useful tool for validation, storage and translation of the semantically annotated documents.

## 7 Conclusions and Future Work

This work considers the problem of providing a tool support for software engineers in ensuring compliance of new and legacy software with the national regulations as well as the legislative documents of the European community.

We extended and generalized the Cerno-based regulation analysis process for a wide range of legal documents, which now includes not only regulations but also directives. More specifically, we considered two European privacy directives and provided a comprehensive evaluation of the proposed process. In addition, we developed a database backend to Cerno, in this way facilitating storage, checking and querying of semi-formal rights and obligations.

The results of this work can be reused for modeling privacy requirements in all domains that develop software systems which should comply with the European privacy directives.

As the direction for future work, we propose to combine the set of rights and obligations generated from the European directives with those provided by national laws, that describe precise implementation solutions chosen by Member State to address the requirements of the EU.

**Acknowledgments.** This work has been funded by the EU Commission through the SERENITY project.

## References

1. Agnoloni, T., Francesconi, E., Spinosa, P.: XmLegesEditor: an OpenSource Visual XML Editor for supporting Legal National Standards. In: Proc. of 5th Legisla-

tive XML Workshop, European University Institute, Fiesole, pp. 239-252. Firenze, European Press Academic Publishing (2007)
2. Antón, A. I.: Goal-based requirements analysis. In: Proc. 2nd IEEE International Conference on Requirements Engineering (ICRE'96), Colorado Springs, Colorado, pp. 136-144 (1996)
3. Antón, A. I., Earp, J. B., He, Q., Stufflebeam, W., Bolchini, D., Jensen, C.: Financial privacy policies and the need for standardization. IEEE Security and Privacy, vol. 2(2), pp. 36-45 (2004)
4. Antoniou, G., Billington, D., Maher, M. J.: On the Analysis of Regulations using Defeasible Rules. In Proc. of 32nd Annual Hawaii International Conference on System Sciences (HICSS'99), vol. 6, pp. 6033-6039, Washington, DC, USA. IEEE Computer Society (1999)
5. Atre, S.: Relational Views of XML for the Semantic Web. Master Thesis. Queen's University, Kingston, ON, Canada (2007)
6. Baeza-Yates R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
7. Berry, D., Kamsties, E., Krieger, M. M.: From Contract Drafting to Software Specification: Linguistic Sources of AmbiguityA Handbook. Technical Report, School of Computer Science, University of Waterloo, Waterloo, ON, Canada (2003)
8. Breaux, T. D., Vail, M. W., Antón, A. I.: Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. In: Proc. of 14th IEEE International Requirements Engineering Conference (RE'06), Minneapolis, St. Paul, Minnesota, pp. 46-55, Washington, DC, USA, IEEE Computer Society (2006)
9. Cover, T. M., Thomas, J.A.: Elements of Information Theory. Wiley & Sons, New York, NY, USA (1991)
10. Dean, T.R., Cordy, J.R., Schneider, K.A., Malton, A.J.: Using design recovery techniques to transform legacy systems. In: Proc. of ICSM'01, pp. 622–631 (2001)
11. Hearst, M.: Automated Discovery of WordNet Relations. In: WordNet: An Electronic Lexical Database, Christiane Fellbaum (ed.) MIT Press (1998)
12. Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J.R., Mylopoulos, J.: Applying software analysis technology to lightweight semantic markup of document text. In: Proc. of 3rd International Conference on Advances in Pattern Recognition (ICAPR'05), Bath, UK, vol. 3686 of LNCS, pp. 590-600. Springer (2005)
13. Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J. R., Mylopoulos, J.: Text mining through semi automatic semantic annotation. In: Proc. of 6th International Conference on Practical Aspects of Knowledge Management (PAKM'06), vol. 4333 of LNCS, pp. 143-154. Springer-Verlag (2006)
14. Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J. R., Mylopoulos, J.: Annotating Accommodation Advertisements using CERNO. In: Information and Communication Technologies in Tourism 2007: Proceedings of the International Conference (ENTER 2007), Ljubljana, Slovenia, 2007, pp. 389-400. Wien: Springer Verlag (2007)
15. Kiyavitskaya, N., Zeni, N., Mich, L., Breaux, T. D., Antón, A. I., Mylopoulos J.: Extracting Rights and Obligations from Regulations: Towards a Tool-Supported Process. In: Proc. of 22nd IEEE/ACM international conference on Automated software engineering (ASE'07), pp. 429-432. ACM Press, New York, NY, USA (2007)
16. Nute, D.: Defeasible reasoning. In Proc. 20th Hawaii International Conference on Systems Science, pp. 470-477. IEEE Press (1987)

17. Palmirani M., Brighi, R.: Norma-System: A Legal Document System for Managing Consolidated Acts. In: Database and Expert Systems Applications: Proc. of 13th International Conference DEXA 2002, Aixen-Provence, France, pp. 295-314 (2002)
18. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In Proc. of International Conference on New Methods in Language Processing, Manchester, UK (1994)
19. U.S. Goverment: Standards for privacy of individually identifiable health information, 45 CFR part 160, Part 164 subpart E. In Federal Register, 68(34), pp. 8334-8381 (2003)
20. van de Ven, S., Hoekstra, R., Winkels, R.: MetaVex: Regulation Drafting meets the Semantic Web. In: Proc. of the Workshop on Semantic Web technology for Law (SW4Law 2007) (2007)