Ontology Enrichment Based on the Mapping of Knowledge Resources for Data Privacy Management

Fernando M.B.M. Castilho[1], Roger L. Granada[1], Renata Vieira[1], Tomas Sander[2], Prasad Rao[2]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Ipiranga Av., 6681. FACIN. CEP 90169-900. Porto Alegre, Brazil.

[2]Hewlett-Packard (HP)
Ipiranga Av., 6681. Building 91B. CEP 91530-000. Porto Alegre, Brazil.

{fernando.castilho, roger.granada}@acad.pucrs.br,
renata.vieira@pucrs.br, tomas.sander@hp.com, prasad.rao@hp.com

Abstract. *This paper presents a mapping of enriched knowledge resources for data privacy management. An ontology, enriched through natural language techniques, is used for an integrated visualization for global inspection of heterogeneous data. The visualization helps stakeholders in exploring and maintaining a knowledge base for data privacy accountability. The integration of resources on the basis of concepts described in an enriched ontology is an aid to Knowledge Management (KM) in a dynamic domain, due to changes in laws and the corresponding system requirements.*

1. Introduction

The use of ontologies helps to achieve consensus on terms related to specialized domains. The mapping of heterogeneous resources from knowledge rich systems can help domain stakeholders in achieving their knowledge intensive related tasks. This paper is contextualized in the data privacy domain, especially considering the task of accountability. One of the main concerns in data privacy accountability is to avoid data misuse in collecting and handling Personal Identifiable Information (PII) [1]. To ensure that an organization needs robust mechanisms to implement its privacy policies.

Weitzner [2] defines: "Information accountability means that information usage should be transparent so it is possible to determine whether a use is appropriate under a given set of rules". One aspect of determining such usage is the identification of privacy risks related to sensitive information[1]. We discuss the integration of knowledge resources of a rule based tool that provides guidance and privacy assessment of a project that handles PII and identifies possible privacy risks. From now on we simply call it 'accountability tool'. Its resources comprise a questionnaire, a glossary of privacy terms, a set of encoded rules, company policies and a set of guidelines for developers.

The motivation of our work relies on the enrichment of an ontology, based on linguistic and knowledge resources in the privacy domain. We developed a visualization tool to integrate these resources. The mapping of knowledge sources and artifacts, based on an ontology model, can provide a better overview of the information handled by the

[1] "Sensitive information" as defined in: TCSEC - Department of defense trusted computer system evaluation criteria. Dept. of defense standard, Department of Defense, Dec 1985.

accountability tool, and thereby support various critical tasks to reduce oversights and errors in the management of privacy in company projects.

The domain stakeholders are privacy officers, Knowledge Base (KB) engineers, and project managers. Privacy officers are generally accountable for compliance with privacy regulation, and for creating, maintaining and checking the correctness of the underlying KB, as well as for evaluating impacts of changes in the body of laws and documents such as company privacy guidelines. They are in charge of transforming laws and regulations into specific company policies and guidelines. KB engineers are in charge of modeling legal constraints and requirements involving policies and laws, and writing and updating rules in the accountability tool. Finally, project managers are responsible for company projects and their alignment with organizational policies. As an example of benefits of the KM, richer information may help project managers to take information into account such as privacy lawsuits in progress, upcoming changes in laws and regulations etc. that are otherwise unlikely to be available to them.

To help stakeholders with their tasks we propose a mapping between various knowledge sources. The existing sources comprise privacy regulation documents, along with the KB of the rule based system mentioned earlier. Privacy documents are regulatory texts like acts, norms and guidelines for privacy assurance and safe, and accountable software development[2]. A domain ontology was developed, which is at the core of the mapping structure. It is enriched by automatically generated resources: a thesaurus and a list of Named Entities (NE) referring to normative regulations in the privacy domain. The idea is that the enriched ontology can serve to maintain the rule based system. Our work is then based on the definition of an automatically enriched conceptual structure, and on the mapping of knowledge sources to ontology concepts, aiming at the establishment of a KB management infrastructure.

This paper is organized as follows: Section 2 introduces related work, with an analysis of the privacy risk management problem and the use of ontologies in this area, and describes the contribution of our model for the representation of data privacy risks. Section 3 presents an overview of linguistic and semantic resources and their integrated visualization based on the enriched ontology. In Section 4 the overall process of integration of the resources is presented, along with the evaluation of the ontology enriching methods for Thesaurus and NE. Finally, in section 5 we present our concluding remarks.

2. Related work

The development process in Information Technology (IT) is one of the main areas on which privacy strategy needs to focus. Taking up to date privacy legislation into account is an important requirement for IT projects. The knowledge of weaknesses in projects with respect to privacy laws and guidelines, as well as their correct application in such projects, can help to correct inadequate procedures or prevent serious privacy incidents such as data breaches, and thus avoid lawsuits and the loss of consumer trust for a company [3].

[2] Asia-Pacific Economic Cooperation. 2004. APEC Privacy Framework. 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, Official Journal L No. 201, 31.07.2002. Microsoft Privacy Guidelines for Developing Software Products and Services, v. 3, 2009. The U.S. Children's Online Privacy Protection Act of 1998 (COPPA). The U.S. Computer Fraud and Abuse Act. U.S. Federal Trade Commission (2000). Financial Privacy Rule.

Our work aims at the identification of privacy risks, and the management of the domain KB that supports it. An ontology was modeled considering the movement of information across borders, and the actions performed on it to identify privacy risks, and to organize thinking and discussion in the privacy field, which is relevant to IT. This approach follows the need stated by Solove [4] who developed a taxonomy of privacy to describe concepts of information collection, processing, dissemination and invasion to capture violations of privacy. It focuses on actions and focuses on activities such as the collection, processing and dissemination of information, which remove it further from the direct control of the user.

A process to promote privacy assurance inside organizations and to establish proper privacy management to address legislative requirements, policy guidance and business standards is proposed in [5]. Knutson [6] presents some principles that organizations should follow to create privacy awareness. He points out that a privacy core team with technical and legal experts must define a privacy terminology to achieve a common understanding of the scope and meaning of rules. Another recommendation is to create guidelines to help developers to become independent from privacy experts with respect to basic tasks. Similar concerns for software design are endorsed within other works on privacy awareness [7][8]. In our work these requirements are carried out with the definition of an ontology enriched by integrated knowledge resources.

Recognizing concepts and instances in text in order to support ontology maintenance and semantically represent the meaning of sentences is a task explored in [9]. One step towards a better control of the development process from a privacy perspective is to have a proper representation of the relevant rules that have already been formulated for handling PII. These rules are mostly described in laws, policies and other normative sources, such as implementation guidelines, best practices and information security standards. There is a rich literature describing ontologies to represent such rules for the security and privacy management area. Abou-Tair [10] presents a way to enforce privacy in enterprises using ontologies to generate XACML [11] policies. The work presents the BDSG (Federal Legislation on Data Protection) ontology in F-Logic mapping law statements to a machine interpretable language. In our work the integration of resources to support the maintenance of a KB on the privacy domain establishes a space for common understanding necessary for the implementation of privacy rules in accordance with legal constraints and local policies among others.

Hecker [12] argues that privacy ontologies must show different concepts and associations, enabling interoperability and determining the privacy level of a transaction. Ontologies can also guide system developers who need to implement privacy functionalities or mechanisms without requiring expertise from developers specialized in the privacy domain. The proposed integration of resources on the basis of an ontology aims at the integration between system developers and other stakeholders in the requirements elicitation task.

Hu [13] proposes that the semantic model for EPAL privacy policies [14] can be expressed as a variety of ontologies and rule combinations. It supports the idea that ontologies are the main body of concepts to establish an infrastructure for the knowledge management in a domain. Our work does not focus on rules. Instead, ontology concepts are mapped to resources in the domain to support the challenge of semantic representation. It defines the basis for the enforcement of privacy, as well as for knowledge management in the privacy domain.

Although there are many ontologies in the privacy domain, reusing them is a difficult task, as they are developed for a wide variety of purposes, which differ from the specifics of our context. Our privacy ontology was manually built, based on the study of regulatory documents, guidelines, and also on some aspects of the KB system. Ontology concepts are then a central structure, from which other knowledge and linguistic resources are generated (Noun Phrase taxonomy, thesaurus, and NE). Several domain documents are mapped to the enriched ontology. Identified concepts and their extensions are then linked to all the domain resources in which they occur. Therefore the privacy ontology serves as a guide to several knowledge related tasks in which domain stakeholders are involved.

3. Knowledge resources

A manually built privacy ontology, validated by a privacy officer, a lawyer, and a project manager, was enriched with other resources on the basis of corpus processing. These resources are composed of a thesaurus, a noun phrase taxonomy, and NE. The corpus-based thesaurus relates terms that are similar to each ontology concept, and constitutes an extra semantic resource for assisting stakeholders. NE guide the access to important law documents. The taxonomy shows concepts related noun phrases organized in a hierarchy, which helps gathering information about contextualization of terms.

The remaining resources support the inference of risks by the accountability tool, comprising a glossary that describes important terminology, a questionnaire and a rule set that guides the flow of questions for privacy assessment of projects, resulting in the inference of the project risk level. Relating the system KB to the enriched ontology and the corpus is considered as an aid for engineers responsible for system updates.

The domain knowledge resources can be accessed on the basis of a given term, selected in a visualization tool available at http://www.cpca.pucrs.br/VisualizationTool/. The Ontology is viewed as a hyperbolic tree of concepts, instances and related properties. Such a view of fundamental domain concepts is then integrated with all the other knowledge resources. Users can then access related concepts in the thesaurus, and from it, navigating through all the accountability tool resources. A resource can be accessed through its tab or through context menus. The following section explains in more detail each accountability tool resource.

3.1 Accountability Tool Resources

This section describes resources directly related to the accountability tool. To accomplish with project restrictions, these resources are presented as figures on the text, and omitted from the visualization tool available.

3.1.1 Glossary

The domain glossary can help clarifying terms to stakeholders, represented as an entry to which a description is given, or as part of the description. An occurrence of *personal information* can be shown in Figure 1.

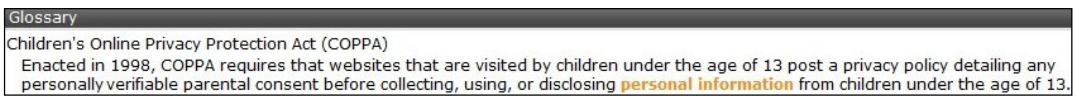

Figure 1 – Excerpt of the Glossary, with "personal information" as part of description

3.1.2 Inference Rules

Inference rules are managed by a risk inference component in the accountability tool. They guide the flow of questions that are shown, as well as determine the project privacy risks based on the answers provided by the user. They are structured as follows: rule name, risk indicator, origin of the rule, reason for the rule, remediation, and condition to fire the rule. As seen in Figure 2, when modeling requirements, KB engineers may learn that a notice statement must be provided by the system before collecting personal information, and also that other issues are involved. Similarly, privacy officers can check which rules will be affected when a change in the body of laws involving personal information occurs. Project managers can access occurrences of the term in the rulebook to help mitigate privacy risks for their projects as well as to manage organizational resources affected by rules related to personal information.

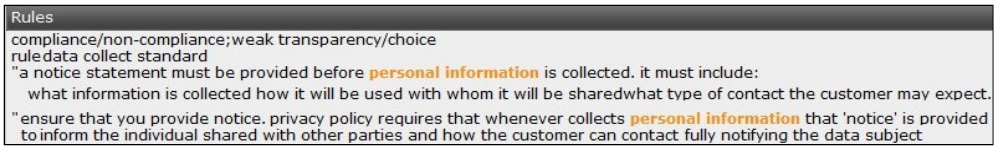

Figure 2 – Excerpt of the rules with the personal information concept

3.1.3 Questionnaire

The questionnaire is managed by the risk inference tool through questionnaire and compliance rules. The former rules involve pairs of questions and their possible answers, and also allow for more user-friendly grouping and ordering of questions. They are also used to set the value of intermediate variables, to decide which questions should be shown to the user, given the answers already provided. Intermediate variables are kinds of flags with a semantic meaning, hand-created by privacy experts to simplify the authoring of the rule base and manage the relation between the data comprised in questionnaire sections, and the knowledge it represents. Based on given answers a set of compliance rules with the form "*when condition then action*" infers the privacy compliance level of the project [15]. A compliance report is generated, with the results of the assessment of privacy risks and a list of remediations in case of higher-risk privacy concerns.

Figure 3 presents the term *personal information* in the question 66. For each question or answer with at least one occurrence of the mapped term the system presents the text of question and answer to better contextualize it. KB engineers can thus gain immediate and comprehensive control of impacts of domain changes to the questionnaire, and along with privacy officers keep the various objects in the rule base aligned with regulations and internally consistent during KB management.

Figure 3 – Excerpt of questionnaire with the personal information concept

The next sections describe the ontology and the remaining corpus-based resources.

3.2 Privacy Ontology

Despite the maturity in this field [16] reuse is difficult since each proposal is created for different purposes. Privacy risk assessment and analysis vary according to the

requirements imposed by specific scenarios. The definition of our Privacy ontology involves modeling concepts from several knowledge sources related to the problem of data privacy accountability, such as a set of legal documents, an accountability tool, and particular rules considered in the privacy risks inference scenario. The overall goal is the reduction of the difficulty of KB maintainability. To better understand the following explanation of the main concepts, we suggest the exploitation of the ontology through the visualization tool. Ontology concepts are represented as seeds on the Thesaurus tab.

Some concepts were chosen to identify references to legal documents. Thus, regulations are classified as normative and non-normative. Regarding the accountability tool, the ontology includes concepts related to project activities and purposes, user information, and sensitive information. Other essential concepts are *PII* and *Sensitive_PII*. People and organizations are also important concepts, because they refer to those involved in a transaction handling PII.

Concerning the idea of privacy risks, the ontology includes different risk levels. *Actions* conducted in a project can be associated to different *Risk Levels*. *Actions* and *activities* with no associated risks are evaluated to a *green level*. When internal policies are violated, the risk level associated with the activity is evaluated to a *yellow level*, and finally, the *red level* is attributed to activities that violate laws or regulations. Geographic locations, classified by the concept *Geo* as *cities*, *continents* and *countries* directly affect the definition of privacy risks.

All these concepts can be used in the description of project actions and their associated risks. In case of transborder data flows, for instance, risks depend on the kind of information, and on the origin and destination of the data flow.

The EU Data Protection Directive 95/46/EC, for example, imposes restrictions on the flow of PII to a third country, outside the European Economic Area [17]. A country is considered adequate for the flow of personal data if its laws provide a level of protection for personal data comparable to the Directive. Otherwise, it is considered non-adequate.

3.3 Corpus

The corpus used in our project was composed of a set of 100 documents of privacy regulations and development guidelines. By accessing the concept *personal information* in the visualization of the corpus, each occurrence of the term in a document is displayed in a context defined by five words on its left and right, which is called a concordance, along with the document identification, and the line number. This link can be used by Privacy Officers to evaluate how these concepts are used in regulations contained in the corpus, e.g., to verify that company practices (or the KB) are aligned with these regulations including in the presence of regulatory changes.

When a user selects the document name in the column Corpus, the original text file is highlighted in a concordance. A KB engineer can have a better understanding of requirements involving the flow of personal information to avoid the transfer of information without some adequate level of protection corresponding to Section 12, Item 1 of the highlighted text, for example. Similarly, a project manager can browse through the corpus of laws and guidelines to discover which documents can affect a system update involving, for example, additional transborder data flows of personal information. Stakeholders can inspect the KB, in order to decide the implications of changes to their respective fields in the organization.

KB engineers can more effectively model requirements involving rules and laws or check for KB correctness through being aided in the interaction with privacy officers by searches in the corpus. The inspection of other resources also helps to clarify a term in the domain, and to become aware of the impact of lawsuits arising from the misuse of personal information among others as seen in the following sections.

3.4 Corpus based Ontology Enriching Resources

The next subsections describe the resources extracted from the corpus directly related to domain concepts.

3.4.1 Thesaurus

As legal documents have large quantities of domain specific terms whose meaning can be represented with different terms the creation and maintenance of a thesaurus is a task that requires technological support. A thesaurus is composed of terms called seeds, to which similar terms in the domain are related. Associating a thesaurus to an ontology, and to a domain corpus can increase the efficiency of document retrieval. Instead of retrieving only documents containing specific terms the ones with terms semantically related can be retrieved. For example considering the term *personal_information* it is also referred to in the corpus as *personal_identifiable_information*, *personal_data*, and as the acronym *pii*. Thus by associating a thesaurus to our privacy ontology instead of retrieving only documents that contain the occurrence of some specific term documents containing also related terms can be the retrieved enriching the results with semantic privacy meaning.

Each ontology concept represents a seed in the thesaurus. To each seed shown on the tab "concept" of the visualization tool, a list of related terms from the corpus on the right was automatically generated using linguistic and statistical techniques. The ontology concept *personal_information* is found as similar to the terms, *PII*, *patient record* and *sensitive information*. By choosing a term in the thesaurus its occurrences in other knowledge resources can be accessed by stakeholders.

3.4.2 Named Entities (NE)

NE can be used to populate an ontology with instances extracted from the domain terms. The automatic recognition of NE from legal and normative documents can help the construction of a conceptual base of the privacy domain. In our work NE from legal texts representing instances of classes that contain as keywords the terms *act*, *law*, and *rule* were used to populate the ontology [18]. A list of classes extracted from the corpus of laws is shown in Table 1.

Table 1 - Examples of classes extracted with NER

Original classes	Derived Classes
Act	Enactment, Number, Turn, Routine, Deed, Bit
Law	Police, Jurisprudence, Constabulary
Rule	Ruler, Normal, Pattern, Prescript, Regulation, Principle, Convention, Formula, Dominion

In the visualization tool the term *personal information* can be viewed along with some recognized NE with the class to which they belong (Act), and the name of the legal instrument that contains each term.

When privacy officers and KB engineers are involved in clarifying legal implications that may affect the definition of requirements involving the protection of

personal information, for example, the identification of NE related to the selected term can provide a list of legal regulations to be investigated. Also the NE classes which are represented by ontology concepts can be identified helping the investigation of conceptual constraints in modeling decisions. Project managers can investigate relations between *personal information* and the laws relating to it, through references to legal documents retrieved on the basis of the term to evaluate possible legal implications of this term on project risks, for example.

3.4.3 Taxonomy

A hierarchy of noun phrases related to domain concepts may help stakeholders with a broader view of the context in which ontology concepts occur in documents. This involves more complex structures, which are not modeled as ontology concepts or instances. The taxonomy can help KB engineers to understand the uses of ontology concepts, thus providing extra information about the domain through the inspection of the contexts in which the term occurs. Our taxonomy was developed by parsing the corpus to extract noun phrase hierarchies. This extraction is based on the identification of noun phrases, and for each one on the identification of its constituents and nucleus. Noun phrases with the same nucleus were grouped and organized in a hyperbolic tree according to its constituents.

The resources described up to now summarize the mapping of textual sources, on the basis of a term for the information handled by domain stakeholders. The following section explores the evaluation performed by domain experts in the thesaurus and in the NE, to validate our efforts aimed towards establishing an infrastructure to the KM in this domain.

Apart from the taxonomy, the previous resources serve as the basis for the annotation of all the other resources for the support of the KM conducted by domain stakeholders as following described.

4. Integration of Heterogeneous Knowledge Resources

The resources mapping task consists in the XML-based indexing of terms from the enriched ontology that occur in a set of domain related documents. The whole process includes the generation of new resources and the annotation of documents in a mapping model shown in Figure 4.

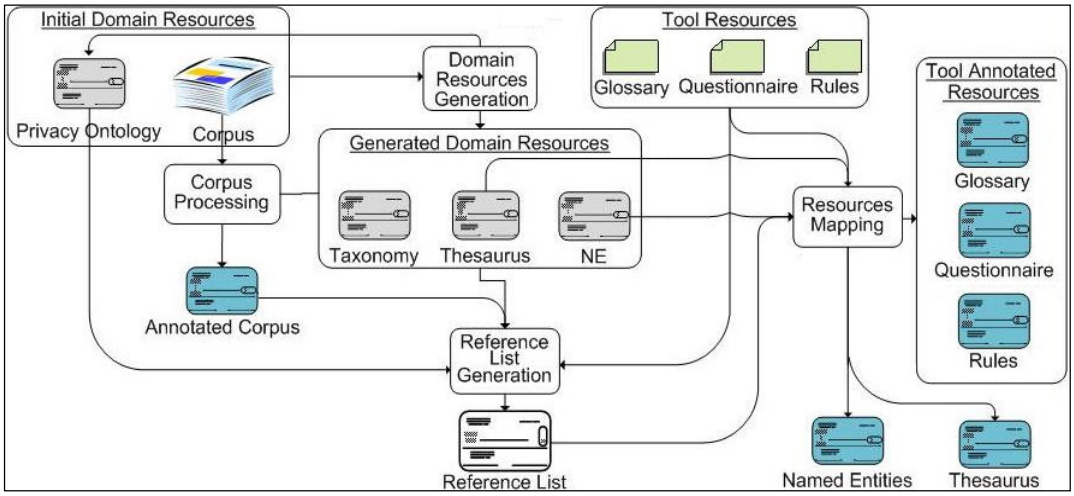

Figure 4 - Resources Generation and Mapping

The mapping process comprises the following steps:

- Generating thesaurus, taxonomy and NE from the ontology and the domain corpus.
- Generating a reference list of terms by merging the domain resources and the ontology concepts and instances.
- Checking the frequency of terms in each knowledge resource for the generation of the reference file.
- Annotating/indexing each knowledge resource based on the reference list.

The mapping procedures must always be performed when the domain KB is updated to set up new relations between resources. The indexing of the term *personal_information* can be seen in Figure 5 which shows the reference list of terms with the presence of the term in each resource.

```
<!-- reference.xml file>
<terms>
    . . .
    <term id="1124">
        <name><![CDATA[personal_data]]></name>
        <resource id="1" name="ontology" term_occurs="false" frequency="0"/>
        <resource id="2" name="thesaurus" term_occurs="true" frequency="1"/>
        <resource id="3" name="corpus" term_occurs="true" frequency="1349"/>
        <resource id="4" name="ne" term_occurs="false" frequency="0"/>
        <resource id="5" name="questionnaire" term_occurs="true" frequency="1"/>
        <resource id="6" name="glossary" term_occurs="true" frequency="1"/>
        <resource id="7" name="tagging" term_occurs="false" frequency="0"/>
        <resource id="8" name="rules" term_occurs="false" frequency="0"/>
```

Figure 5 – Excerpt of the reference XML file

All the other resources are represented by their specific XML files relating the presence of the term in it by its identifier, and specific attributes like the document number and line in which it occurs in the corpus.

4.1 Evaluation of the Ontology Enriching Methods

The evaluation process in our work consisted in verifying the quality of thesaurus generation and precision, recall and coreference for the NE recognition. Although tests for the evaluation of the overall integration of resources were not performed we performed the evaluation of thesaurus and NE, the most important resources that directly affect the enrichment of the domain ontology.

The evaluation of the thesaurus generation was performed by domain specialists including a privacy officer, a lawyer, and a project manager for a sample containing 10 domain concepts and 90 similar terms. The chosen concepts were: *children*, *consent*, *customer*, *data_protection*, *data_subject*, *marketing*, *notice*, *personal_data*, *personal_information*, and *regulation*. To evaluate them, specialists could assign a term as "*similar*", "*not similar*", or "*not sure*" (about the similarity). A term can also be ranked through arrows changing its position in the similarity list. A higher position on the list indicates a higher similarity level.

The precision rate for the sample of similar terms in the evaluation was 51.1%. We cannot fairly compare our results with related work because we do not share the same data. In practical terms, the production of a list of related terms in which about half is likely to be considered useful (as in the case of our methods over our corpus) is an important aid for the knowledge engineering processing.

For the NE recognition three classes of *Normative_Regulation* were considered, namely *Act*, *Law* and *Rule*. Other classes were generated from them, as follows: *Act* (*Enactment*, *Number*), *Law* (*Police*, *Constabulary*) and *Rule* (*Prescript*, *Regulation*,

Principle, and *Convention*). For instance, the class *Number* resulted from the NE *New Tax System (Australian Business Number) Act 1999.*

The Privacy corpus was tagged for these NE. The tagging task resulted in 4863 references to NE and 1191 unique entities in the domain. An evaluation tool analyses the tagging output against a manually tagged reference to obtain precision, recall and F-measure for:

a) Unique entities, represented by unique references to entities names;
b) Repeated references to the same entities.

The evaluation of the NE recognition performed on the corpus found 389 out of 1191 unique entities. An amount of 1460 references out of 4863 were found [18]. Resulting measures including precision, recall and F-measure are presented in Table 2. The results were considered promising and comparable to the results obtained from the 2008 ACE Local ERD. However, the application of more sophisticated natural language processing techniques over larger corpora can improve our results, in particular the recall measure [18].

Table 2 - NER processing resulting measures

	Precision	Recall	F-Measure
References to entities	60.48% (1460 / 2414)	30.02 % (1460 / 4863)	40.13
Unique entities	40.06 % (389 / 971)	32.66 % (389 / 1191)	35.99

We also evaluated coreference (or "*same_as*" relations), based on the search of acronyms. Table 3 has a row that represents both "*Employee Retirement Income Security Act of 1974*" and "*ERISA*", as they were found in the corpus as legal NE, and the system identified them as referring to the same entity. *ERISA* is said to be an acronym of "*Employee Retirement Income Security Act of 1974*". The evaluator was supposed to determine if this relation is correct or not, for the 185 instances related to it.

Table 3 – Acronyms for the relation "*same_as*"

Class	NE	Relation	Class	NE
Act	Employee Retirement Income Security Act of 1974	same_as	Act	ERISA
Act	TCPA	same_as	Act	Town and Country Planning Act 1990
Act	TCPA	same_as	Act	Telephone Consumer Protection Act of 1991

The evaluation of the relation "*same_as*" is presented in Table 4 according to 2 evaluators. We believe that the extraction of semantic relations between the entities recognized in this work and those which relate region-specific laws to their specific geo-political units can improve these results [18].

Table 4 - Evaluator's results for the relation "*same_as*"

	Evaluator 1	Evaluator 2
Correct	52.97 % (98 / 185)	67.03 % (124 / 185)
Incorrect	47.03 % (86 / 185)	32.97 % (60 / 185)

Concerning the taxonomy an evaluation was not performed since the resulting structure is just a straightforward reorganization of syntactic structures. However, the taxonomy generation tool [19] was previously evaluated in [20].

The enriching techniques developed so far can be considered as semi-automatic processes, whose output must be checked by experts given that the error rates are still considerably high. Suggestions of terms are provided by these techniques but an expert is needed in order to approve or refuse these suggestions. However these areas of NLP

are still under development and it is likely that the near future will bring new techniques with better recall and precision.

5. Concluding Remarks

Our work describes an ontology-based integration of knowledge resources in the privacy domain to support an accountability tool, focusing on the definition of concepts and the automatic enrichment of a privacy ontology, and on the construction and mapping of knowledge resources to support KM in the domain. The generation of relations between ontology concepts and various knowledge sources established the basis for knowledge inspection and refinement in accordance with changes in laws or in policies and requirements of the organization. The impact of such changes on the resources can be evaluated with the help of the integrated visualization tool developed in the project.

The domain concepts defined in the privacy ontology can be used to support the maintenance of the accountability tool. Our efforts were aimed at the definition of the mapping structure to integrate domain resources, and at the deployment of a tool to permit stakeholders to explore the knowledge and evaluate impacts of changes in the domain. As a result our ontology is composed of 113 concepts and 268 instances.

These efforts resulted in a semantic support that can help navigate through several resources and documents. The generated thesaurus can help specialists to identify similar terms for information search. NE are useful to keep track of changes in laws that need to be considered for KB maintenance. The integrated visualization of knowledge resources can help finding terms in a vast corpus of laws and other domain documents on the basis of an enriched ontology.

The Privacy ontology could not be fully reused for the management of privacy in different companies because it was defined to support a specific accountability tool, and it refers to concepts and instances of Project Activity and User modeled according to specific requirements. However, most concepts will remain useful in an ontology engineering process for similar ontologies.

We consider exploring more specialized semantic relations and features for the automatic recovery of information as future work.

References

1. Yee, G.O.M.; Korba, L.; Song, R. (2008) “Cooperative Visualization of Privacy Risks”, In: 5th International Conference in Cooperative Design, Visualization and Engineering, LNCS, vol. 5220, pp. 45-53.
2. Weitzner, D. J.; Abelson, H.; Berners-Lee, T.; Feigenbaum, J.; Hendler, J.; Sussman, G. J. (2008) “Information accountability”, *Commun. ACM* vol. 51, no. 6, pp. 82-87.
3. Mont, M.; Thyne, R. (2006) “Privacy policy enforcement in enterprises with identity management solutions”, In: PST '06, vol. 380, pp. 1-12.
4. Solove, D. J. (2006) “A Taxonomy of Privacy”, *University of Pennsylvania Law Review*, vol.154, no. 3, p. 477.
5. Rachamadugu, V.; Anderson, J. A. (2008) “Managing Security and Privacy Integration across Enterprise Business Process and Infrastructure”, In: 2008 IEEE Intl. Conf. Services Computing, vol. 2, pp. 351-358.

6. Knutson, T. R. (2007) “Building Privacy into Software Products and Services”, *IEEE Security and Privacy*, vol. 5, no. 3, pp. 72-74.

7. Duncan, G. (2007) “Engineering: Privacy by Design”, *Science*, vol. 317, n. 5842, pp. 1178-1179.

8. Ye, X.; Zhu, Z.; Peng, Y.; Xie, F. (2009) “Privacy Aware Engineering: A Case Study”, *Journal of Software*, vol. 4, no. 3, pp. 218-225.

9. Schäfer, U. (2006) “OntoNERdIE-Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources”, Proceedings of the 5th International Conference on Language Resources and Evaluation LREC - Volume 07.

10. Abou-Tair, D.D.; Berlik, S.; Kelter, U. (2007) “Enforcing Privacy by Means of an Ontology Driven XACML Framework”, In: IAS 2007, Third International Symposium on Information Assurance and Security, pp. 279-284.

11. OASIS XACML Technical Committee. (2003) “eXtensible Access Control Markup Language”.

12. Hecker, M.; Dillon, T. S.; Chang, E. (2008) “Privacy Ontology Support for E-Commerce”, *IEEE Internet Computing*, vol. 12, no. 2, pp. 54-61.

13. Hu, Y.; Guo, H.; Lin, A. G. (2008) “Semantic Enforcement of Privacy Protection Policies via the Combination of Ontologies and Rules”. In: IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pp. 400-407.

14. Backes, M.; Pfitzmann, B.; Schunter, M. (2003) “A toolkit for managing enterprise privacy policies”, In: ESORICS’03, vol. 2808, pp. 162-180.

15. Pearson, S.; Rao, P.; Sander, T.; Parry, A.; Paull, A.; Patruni, S.; Dandamudi-Ratnakar, V.; Sharma, P. (2009) “Scalable, Accountable Privacy Management for Large Organizations”. In EDOCW 2009, pp. 168-175.

16. Cybenko, G. “Why Johnny can't evaluate Security Risk”. *IEEE Security and Privacy*, vol. 4, no. 1 , p. 5.

17. European Commission: Commission decisions on the adequacy of the protection of personal data in third countries. From the Internet, accessed in 11/25/2010 on the URL: http://ec.europa.eu/justice/policies/privacy/thridcountries/index_en.htm

18. Bruckschen, M.; Northfleet, C.; Silva, D. M.; Bridi, P.; Granada, R.; Vieira, R.; Rao, P.; Sander, T. (2010) “Named Entity recognition in the legal domain for ontology population”, In: LREC 2010, pp. 16-21.

19. Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. (2009) “ExATOlp - An automatic tool for term extraction from portuguese language corpora”, In: LTC’09 - 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics.

20. Lopes, L.; Oliveira, L.; Vieira, R. (2010) “Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches”. In: PROPOR 2010 - International Conference on Computational Processing of Portuguese Language.