Ontological support to knowledge management in a hydrogeological information system

M. T. Pazienza, M. Pennacchiotti & A. Stellato
AI Research Group, Department of Computer Science, Systems and Production, University of Roma, Tor Vergata, Italy

Abstract

In this work we report our experience in realizing an Information System for the Italian APAT agency (Azienda per la Protezione Ambiente e Territorio) with the aim of supporting analysis of the hydrogeological situation of the Italian territory. Objective of the system is to provide a structured environment for knowledge management and report production, to support the complex activity of APAT officers in charge of retrieving, organizing and managing data originating from distributed APAT agencies (one for each Italian region) and of those involved in the production of synthesis documentation over the collected information.
Keywords: information systems, ontology management, terminology extraction.

1 Introduction

Management of complex administrative processes, dealing with considerable amount of data to be stored, retrieved and evaluated, is a difficult task, involving lot of technological and human resources. Coordination of the different figures of competency which collaborate at different levels of the workflow, and which actively access to the same knowledge, is thus an important aspect to be dealt with. Identification and implementation of dedicated modalities for accessing and managing information according to different contexts and user competencies is one of the key issues that research in modern Knowledge Management System should address. In this work we report our experience in realizing an Information System for the Italian APAT agency (Azienda per la Protezione Ambiente e Territorio) with the scope of supporting analysis of the hydrogeological situation of the Italian territory. Hydrogeological data is

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060431

collected in a central repository after an heavy unification process performed over contributions from APAT agencies distributed all over the Italian territory, and constitutes the global information for an analysis process activated by national and/or European regulations changing each year. The system must ensure a global and harmonizing view over the same data, being accessible by different user roles and presenting its knowledge content according to the specific competencies and skills which are exhibited by the user. The KM system follows an hybrid approach to knowledge representation, integrating an high level, conceptual perspective of the domain of interest given by a domain ontology, with massive amount of data which is collected in (and retrieved from) a data base repository. Mapping between the database and the ontology is realized through an instance migration engine which processes SQL statements whose parameters are bound to the ontological entities to be populated. With this approach, arbitrary data structures can be extracted from the DB repository and expressed in term of complex ontological constructs. Maintenance of the ontology is supported by a dedicated terminology-extraction module, which is in charge of analyzing documents, such as new law decrees or legislative acts, containing added or modified regulations and principles for the achievement and respect of standards of quality. The extracted terminology may contain very general concepts, which typically refer to entities of the domain (such as "reservoir", "river", "water", "region", "site" etc) as well as specific information pertaining to the objectives of the analysis of the situation of the territory. Finally, automatic production of reports on the hydrogeological condition of the Italian territory is supported by a dedicated report ontology, whose elements describe the objectives and requirements of every reporting document. In the rest of the paper, we first show the overall architecture of the system, we then offer a detailed description of the single components which characterize the developed KM system.

2 System Architecture

The System Architecture *(Figure 1)* is centered about a customized version of the popular ontology editing framework Protégé [6], which embodies, through different perspectives over the same knowledge data, all the specific aspects of knowledge management. The rest of the architecture is characterized by a combination of external tools and of integrated components, mostly developed as Protégé plug-ins. Though an architecture schema does not offer a dynamic representation of a system design, it emerges from the figure an ideal perspective over the natural flow of information. A terminology extraction component is in fact in charge of analyzing domain documentation (legislative acts regulating the analysis of the hydrogeological condition of the Italian territory, as well as technical documentation) to extract terms relevant for the domain of interest. The ontology editing facilities of Protégé have been enhanced to exploit the available domain terminology (together with possible alternative/synonymic linguistic expressions coming from several linguistic resources, which are accessed

through the linguistic enricher component) in order to build the internal domain representation.

A modular approach has also been followed in structuring the knowledge managed by the system. Two separate ontologies, expressed in the OWL Web Ontology Language [5], have been designed: the first one contains an explicit representation of the domain; the second one, the report ontology, contains those concepts which express clear statements and requirements about the hydrogeological situation of the Italian territory. While the former is composed of concepts characterized by a certain "stability" ("river", "lake", "terrain"), the latter expresses a more dynamic perspective over the domain ("contaminated water", "emission limit"). Very specific information is maintained in a database, which regularly feeds ontological knowledge through a dedicated component (see next chapter). A report generator component is then in charge of managing different document templates (which are also part of the report ontology) and of exploiting their content for the production of reports.

3 Knowledge migration between ontology and database

When we started working on this KM system, we had to face the awesome amount of available data and, moreover, the complexity of its associated model, inherited from a previous system realized by APAT. That system, consisting in a traditional DBMS, was characterized by more than 120 tables, whose complexity was very difficult to manage: different aspects like conceptual representation of the domain, individuals, specifications for report documentation, etc all of them were mixed in a single db schema. Furthermore, as a project requirement, we had to make possible a smooth passage from the old system to the new one.

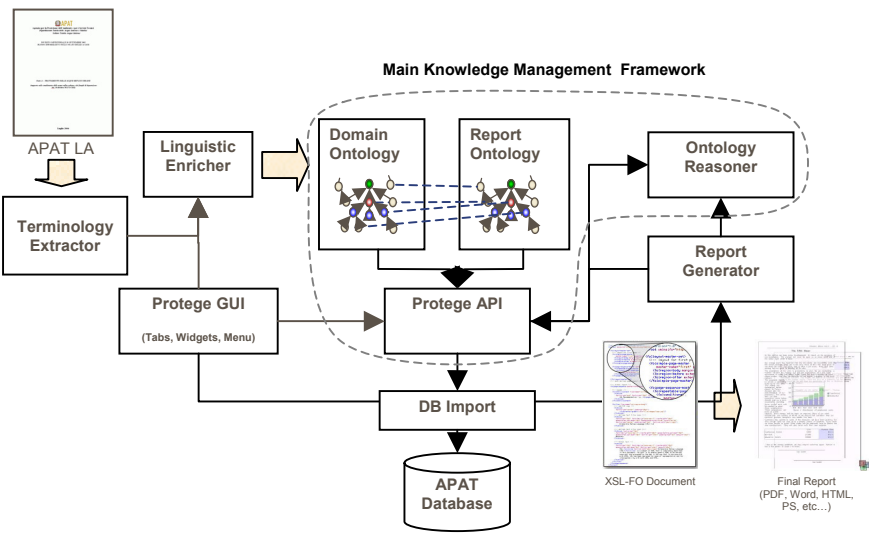

Figure 1: System architecture.

We looked for a hybrid approach in which very specific information (e.g. percentages of each single substance in the water) were left in the original database, while the ontology – and the related knowledge base – was intended to summarize the underlying information with less detail but with greater emphasis on clarity of description. Initially, we started considering existing tools for exchanging data between ontologies and databases, but the only one we found, DataGenie [11], imposed a strict policy which did not admit any choice on the way schema and data information from the database have to be imported. As the DataGenie documentation reports: "Generally, each table becomes a class, and each attribute becomes a slot. In addition, if the relational database table has foreign key references to other tables, these can be replaced by Protégé instance pointers when the database is converted into a knowledge base", it is clear that it is not possible to realize the dynamic import mechanism we needed in our project, in which both the db schema and the ontology schema are already specified and what is needed is to move data from one schema to another according to their specific formats.

We thus developed a dedicated import system, implemented as a Protégé plug-in, which associates an SQL query to each ontological object, enabling its description to be updated by retrieving data from the db. The db_import plugin recognizes two specific properties inside the ontology:

- `use_sql_query`: specifies the query which is used to retrieve data from the database and associates it to a given class (the class the property is attached to). A query may be marked as bound or unbound, in the first case, the query is ground and can be invoked as it is to retrieve data from the db, while if it is unbound, it contains variables which need to be bound to ground values.
- `use_sql_bindings`: this property contains variable-pair pairs which are used to make unbound queries become ground SQL statements.

An expression like the following:

```
SELECT Count(*) AS numero_agglomerati, regione FROM
sch6_1_agglomerati WHERE cariconominale >= %MAXLOAD% and <=
%MINLOAD%
```

can thus be associated to a class which is occasionally restricted on the property `use_sql_bindings` to the values:

```
MINLOAD == 10000 ; MAXLOAD == 20000
```

This way, the same query may be reused to retrieve information for several ontology concepts, by adopting different bindings of its parameters. This approach facilitates reuse of existing SQL code. Once documented inside the ontology and provided with dedicated slot fillers for specifying its unbound parameters, each query may be adopted and customized even by non technical administrative staff.

4 Terminology extraction

4.1 Terminology extraction engine

Terminology extraction and organization have assumed in recent years a central role in new NLP tasks devoted to coherently build and organize domain knowledge. As a matter of fact, Ontology Learning and Linguistic Ontology Enrichment are widely seen as useful means to semi-automatically support the often time consuming and difficult building process of ontologies [1]. In this view, the availability of NLP technologies able to automatically extract domain linguistic knowledge (terms and term relations) are seen as a first step towards the creation of a concept hierarchy and relational network.

In many frameworks such processes start with terminology extraction. Terms are automatically extracted from a domain corpus and then manually validated (and possibly enriched) by a domain expert. Different methodologies have been adopted for extracting terms, ranging from pure statistical techniques to linguistic based approaches. In between, hybrid system mixing linguistic knowledge with corpus statistical evidence has been shown to be both reliable and accurate.

The terminology extraction component developed at the University of Tor Vergata adopts a hybrid approach, realized through a *linguistic module* and a *statistical module*. The *linguistic module* takes as input the corpus from which terms must be extracted and outputs a flat list of *candidate terms*. A candidate term is defined as linguistic form extracted from the corpus that has all the linguistic properties of a term. For example, all the *nouns* of the corpus are candidates, as nouns are one of the most common forms of terms (e.g. *water, policy, quality*). Aim of the *statistical module* is to filter the candidate terms and rank them according to their relevance for the domain, using their frequency in the texts (or other statistical measures) as relevance score. False candidate terms can be then filtered out as unreliable, using a threshold cut on the score.

4.1.1 Linguistic module

The linguistic module extracts candidate terms from the domain corpus using the two sub-modules described below.

- A parsing sub-module that performs a shallow linguistic analysis. The analysis is carried out by the Chaos parser [2], a modular dependency-based syntactic parser for Italian and English texts, developed at our laboratory. Using Part of Speech (PoS) tagging techniques the module identifies *nouns*, *verbs*, *adjectives* and other part of speech in the text. Successively, noun phrase are recognized by a surface syntactic analyzer. Output of the module is a syntactically analyzed corpus;
- A term recogniser sub-module, that using regular expressions extracts from the tagged text only *admissible surface forms* (*candidate terms*), filtering out non interesting forms. These forms represent good syntactic prototypes of candidate terms. *Table 1* reports the *admissible surface forms* used by our system, classified in *k-word* categories, where *k* indicates the number of *main*

items (i.e., meaningful words, as nouns, verbs and adjectives) contained in the term. Moreover, stop lists are used as linguistic filters to discard candidate terms containing common words (e.g. *"this day"*), as they are most likely false terms. The use of stop list including specific type of adjectives and highly frequent nouns in language, have been proved in [9] to be of great help in improving the performance of the system.

The linguistic module outputs a flat list of candidate terms that are used as input to the statistical module.

Table 1: Admissible surface forms used by the linguistic module, coded using regular expressions.

TERM CLASS	SURFACE FORM
1-word	(noun)
2-word	(adj)(noun) (noun)(noun) (noun)(prep)(noun)
3,4,5-word	(noun){3,5} (noun)(prep)(noun){2,4} (adj)(noun){2,4}

4.1.2 Statistical module

The statistical module ranks candidate terms extracted by the linguistic module according to a chosen information theoretic measure. Candidate terms appearing higher in the rank are considered more reliable. Different measures have been proposed in the literature to rank terms [8]. There is still not shared agreement on which measure guarantees the best rank. What is expected is that an ideal statistical measure should be the one that better grasps the definition of term (called *termhood*).

Our extraction system adopts a pragmatic approach to term ranking, allowing the use of different measures. However, [9] and [4] showed that *frequency* is in most cases the best measure to adopt. For the purpose of our application we then use frequency as a ranking measure.

The use of frequency is supported by the simple assumption that a frequent expression denotes an important concept and should then assume a high position in the rank of candidate terms. The most important objection in using frequency is the fact that it doesn't take into consideration the degree of association (*unithood*) among words composing multiword terms [3]. Thus, very frequent expressions are considered good candidates while not being terms (e.g. *"this day"*). In order to capture indirectly the *unithood* nature of terms while using frequency, our system implements linguistic filters, as described in the previous section.

Once ranked, false terms are filtered out using a threshold τ on the frequency score. Experimental evidence in [9] on medium size corpora (as the one at hand) has demonstrated that $\tau=5$ offers a good compromise for obtaining high Precision while not loosing much Recall.

Table 2: Best ranked terms extracted by the terminology extraction system for different term length.

1-WORD TERMS	*FR*	*2-WORD TERMS*	*FR*	*3-WORD TERMS*	*FR*
acqua *(water)*	936	Acqua reflua *(waste water)*	155	acqua_refluo_urbano *(urban waste water)*	57
articolo *(article)*	741	Corpo idrico *(water body)*	138	piano_di_gestione *(management plan)*	56
stato *(status)*	362	Presente direttiva *(actual directive)*	87	gestione_di_bacino *(basin management)*	55
decreto *(decree)*	351	decreto legislativo *(legislative decree)*	86	obiettivo_di_qualita *(quality goal)*	53
scarico *(drainage)*	263	Provincia autonoma *(autonomous district)*	47	data_di_entrata *(entry date)*	41
direttiva *(directive)*	261	Area sensibile *(sensitive area)*	41	trattamento_di_acqua *(water treatment)*	36
legge *(law)*	224	Servizio idirico *(water service)*	41	scarico_di_acqua *(water drainage)*	35
qualità *(quality)*	213	Bacino idrografico *(river basin)*	32	elemento_di_qualita *(quality element)*	31
regione *(region)*	207	Riferimento normativo *(normative references)*	31	stato_di_acqua *(water status)*	31
area *(area)*	184	Sostanza pericolosa *(dangerous substance)*	31	programma_di_misura *(measure program)*	27
sostanza *(substance)*	154	Obiettivo ambientale *(environmental objective)*	28	Tutela_di_acqua *(water protection)*	26
ambiente *(environment)*	141	Risorsa idrica *(water resource)*	28	Piano_di_tutela *(protection plan)*	24
bacino *(basin)*	135	Consumo umano *(human consumption)*	27	standard_di_qualita *(quality standard)*	23
piano *(plan)*	132	Lavoro pubblico *(public work)*	27	decreto_di_ministro *(ministerial decree)*	22
monitoraggio *(monitoring)*	131	Specifica destinazione *(specific destination)*	25	limite_di_emissione *(emission limit)*	22

4.2 The APAT terminology

For the APAT application, the terminology extraction engine has been applied to an Italian corpus of Italian and European Union decrees and legislative acts for supporting the analysis of the hydrogeological situation of the Italian territory. The corpus contains added or modified regulations and principles for the achievement and respect of standards of quality. Four distinct documents form the corpus:

- *Decreto Legislativo 11 maggio 1999, n. 152* (35.544 words)
- *Direttiva 2000/60/CE del Parlamento Europeo e del Consiglio* (22.356 words)
- *Decreto Ministeriale del 18 Settembre: Flusso Informativo sullo stato delle acque* (4.801 words)
- *Decreto Ministeriale 12 Giugno 2003:Regolamento recante norme tecniche per il riutilizzo delle acque reflue* (8.205)

The linguistic module extracted 9.979 terms. The final list of ranked term consists of 917 terms of different lengths, reported in *Table 2*.

Table 3: Examples of wrong terms extracted by the system, together by the related system error.

FALSE TERM	*SYSTEM ERROR*
origine *(origin)*	Common word
termine *(term)*	Common word
fisico-chimica *(physico- chemical)*	Wrong PoS tagging
argomento *(argument)*	Wrong PoS tagging
data *(date)*	Common word

As *Table 2* shows, the extracted terminology reflects the nature of the corpus: knowledge refers both to the generic law domain (e.g., *article, directive, law*) and to the specific domain of hydrogeological analysis (e.g., *basin, water, monitoring*). This is in line with the conceptual knowledge modelled in the Domain Ontology. The linguistic enrichment can be then coherently achieved. The terminology extracted by the system is then ready to be uploaded into the existing Domain Ontology. This process is carried out by a customized version of the *OntoLing* Protégé plug-in [10]. Terms are uploaded in the ontology and bound to concepts as linguistic labels. Terms are then used as a syntactic-semantic interface to find conceptual links between the domain knowledge as represented in the ontology, and the linguistic knowledge, as embodied in the APAT corpus.

4.2.1 Terminology evaluation and analysis

We evaluated the APAT terminology using a *validation approach* (as in [1, 7]) for calculating Precision, and a *reference list* (see [1]) for calculating Recall. As high Precision is a key issue for the overall application, we evaluated the 917 terms extracted by the system with the threshold set to $\tau=5$.

For evaluating Precision, we randomly selected 50 of the 917 terms. A human expert validated each term as valid or not valid for the domain. Precision has then been evaluated as the percentage of valid terms over all extracted terms. To calculate Recall we used as golden standard a term list previously built by APAT experts. We then randomly selected 50 terms from the reference list, and verified if each term appeared in the terminology extracted by the system.

Overall Precision is 70%, while Recall is 43%. These results indicate an overall high quality of the terminology extracted in term of Precision, as needed in the application. While the high cut on frequency ($\tau=5$) significantly affects Recall, this is a minor issue for the application, as less important terms (frequency ≤ 5) are expected to have no reference in the ontology, where only the most important domain concepts are represented.

In *Table 3* examples of wrong terms extracted by the system are reported. As table shows, wrong terms identification is partially due to an erroneous Part of Speech interpretation carried out by the syntactic parser. Other wrong terms are those represented by common nouns that do not fit into the definition of *term* (as a term should be a linguistic expression *specific* for the domain).

In [9] it is shown that the same system applied to a different and more generic corpus achieved a much lower Precision (47%) while maintaining a similar level of Recall. The good performance obtained in the APAT study is mainly due to the focused and well-defined nature of the legislative corpus. Such performance allows a direct uploading of the terminology in the ontology, without the need of a manual validation step. The whole process of terminology extraction can be then intended as being completely unsupervised. The application of the system to new laws and decree on the same domain, as requested by the overall architecture, can be then carried out with minimal human intervention.

5 Report generation

The role of the Report Generation component is to support the APAT officer in the redaction of reports about the hydrogeological status of the Italian territory. This is accomplished by semi-automatic synthesis of human readable documents containing relevant information for the production of new reports. Maximization of Content Reuse is a characterizing feature and a central aspect of this component, which treats documentation as a "first class citizen" in the system ontology, that is, as part of the same knowledge it is meant to describe.

Documents are thus arranged into classes describing those aspects of knowledge which need to be documented. Each document-class finds its counterpart into a Document Template, which represents an empty model for the production of a report. A specific report is thus an instance of a document-class which conforms to its Document Template. The Document Templates contain a mixture of natural language content and semantic pointers to the ontological data stored in the system. When a new report needs to be produced, the user chooses the document-class which best suits its needs and creates an instance of it. At instance creation, a report is generated upon its related Document Template, and all the semantic pointers are resolved by querying the knowledge base. Semantic Pointers take the form of atomic ontological expressions like the following:

```
Instance.Property
```

which returns the value associated to a `Property` in the context of the given `Instance`. More complex queries can be directly expressed in the form of axioms or rules inside the ontology itself, by assigning these expressions to a named class which can in turn be referenced in a Semantic Pointer inside a Document Template. An example is given by the following phrase:

"The maximum depth of the Como lake is $ComoLake.MaxDepth$ meters"

In this case, the Semantic Pointer is resolved by finding the `ComoLake` individual inside the ontology and by reporting the value assumed by that instance on the DataType property `MaxDepth`. Use of variable elements is also admitted, so that, for example:

"The maximum depth of the $?Lake$ lake is $?Lake.MaxDepth$ meters"

In this case, the "?" symbol before the class `Lake` means that the given instance must be chosen by the user at "report generation time". A unification mechanism guarantees that the same instance is chosen one time for all of its occurrences (the scope of a variable is, by default, extended to the whole document).
Two technologies have been experimented and exploited into two different realizations of our report production mechanism:

1. FOP (Formatting Objects Processor) [12]: a technology based on XSL-FO (Extensible Stylesheet language for Formatting Objects), an XML-based markup language promoted by the W3C consortium for describing the formatting of XML data for output to screen, paper or other media.
2. WordprocessingML [13]: a Microsoft technology which provides an XML schema for representing the content of MSWord documents. Every MSWord document can thus be exported in the form of an XML sheet which is validated according to the WordprocessingML schema, and thus be read and manipulated through any XML compliant technology.

In the first realization, document-classes are completed with properties describing the main formatting aspects and the structure of a document (Title, headings, sections, paragraphs, headers etc). This way, it is possible to use property restrictions and subclassing patterns to provide a detailed organization of the kind of documents which may be required. A report is then totally originated from one of these descriptions, by translating the logical structure of the document into an XSL-FO instance and then by producing the physical document via the FOP processor. This approach is focused on maximising content reuse and integration, as each document is not only represented as part of the ontology, but it is considered a non-atomic object. Parts of documents may in fact be shared across different templates and thus reused and integrated with ease, leading to easy composition of those models which best suit user needs. The second realization favours instead ease of use: the user writes a traditional MS Word document, hides ontology queries inside specific fields (the "Quote" field), saves the document in WordprocessingML format and finally associate it to an existing document-class inside the report ontology.

6 Conclusions

It is our opinion that Knowledge Management must go far beyond known aspects like choice of proper knowledge representation languages and/or systems for accessing their content, and meet further high level requirements like supporting users in creating knowledge resources, articulating several perspectives and views over represented data which may vary according to the objective as well as to the different competences and skills exhibited by users. The realized system, with its dual core of ontology/database administration, support for ontology development through extraction and analysis of domain terminology, and automatic production of report documentation, represents a

first insight view over these new possibilities. Though partially immature in its first incarnation, the system demonstrated the potentiality of these new approaches and, most importantly, even at this first stage of development immediately revealed to be a desirable choice for users who have tested it in this first release.

References

[1] Basili R., Pennacchiotti M., Zanzotto F.M.: Language Learning and Ontology Engineering: an Integrated Model for the Semantic Web. 2nd Meaning Workshop, Trento, Italy, February 2005.

[2] Basili, R., Zanzotto, F.M.: Parsing engineering and empirical robustness. Natural Language Engineering 8/2-3 (2002)

[3] Basili R., Missikoff M., Velardi P.: Identification of relevant terms to support the construction of Domain Ontologies, ACL workshop on HLT, Toulouse, France. (2001)

[4] Daille, B. : Approach mixte pour l'extraction de termilogie: statistique lexicale et filters linguistiques. PhD Thesis, C2V, TALANA, Universitè Paris VII (1994)

[5] Dean, M. and Schreiber, G. editors: OWL Web Ontology Language Guide. 2004. W3C Recommendation (10 February 2004).

[6] Gennari, J. , Musen, M., Fergerson, R., Grosso, W., Crubézy, M., Eriksson, H., Noy, N. and Tu, S. The evolution of Protégé-2000: An environment for knowledge-based systems development. International Journal of Human-Computer Studies, 58(1):89–123, 2003.

[7] Jones L.P., Gassie E.W., Radhakrishnan S.: INDEX: The statistical basis for an automatic conceptual phrase-indexing system. Journal of the American Society for Information Science 41(2) (1990) 87-97

[8] Kageura K., Umino B.: Methods of automatic term recognition. Terminology, 3(2). (1996)

[9] Pazienza M.T., Pennacchiotti M., Zanzotto F.M.: Terminology extraction: an analysis of linguistic and statistical approaches. To be published in Knowledge Mining, S.Sirmakessis (Ed.), Series: Studies in Fuzziness and Soft Computing, Springer Verlag, 2005.

[10] Pazienza, M.T. and Stellato, A. An open and scalable framework for enriching ontologies with natural language content. The 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'06), special session on Ontology & Text Annecy, France, June 27-30, 2006

[11] DataGenieTab http://protege.stanford.edu/plugins/datagenie/index.html

[12] FOP http://xmlgraphics.apache.org/fop/

[13] Overview of WordprocessingML, Microsoft corporation, November 2003 http://rep.oio.dk/Microsoft.com/officeschemas/wordprocessingml_article.htm