

A Modular Framework for Ontology-based Representation of Patent Information

Mark Giereth^a Steffen Koch^a Yiannis Kompatsiaris^c Symeon Papadopoulos^b
Emanuele Pianta^d Luciano Serafini^d Leo Wanner^e

^a *Universitaet Stuttgart, Germany*

^b *Aristotle University of Thessaloniki, Greece*

^c *Informatics and Telematics Institute, Greece*

^d *Fondazione Bruno Kessler IRST, Italy*

^e *ICREA and Universitat Pompeu Fabra, Spain*

Abstract. In this paper, we present a new ontology-based formalism for representing patent information. The framework defines concepts and relations for the major aspects of patent information pertaining to patent metadata, patent content structure, patent semantics, and patent classification. Each aspect is logically covered by an individual ontology module. The paper further discusses aspects of ontology learning and integration of non-textual information for the patent domain.

Keywords. Patent Information, Semantics, Ontology Learning

1. Introduction

Patents are a valuable, large scale source for scientific and technological information. They have a high impact on research, development, trade, and society. During the last years there have been great efforts in making patent information available electronically; public services such as *esp@cenet* [2] and the *Open Patent Services (OPS)* [6] are prominent examples of such initiatives. Today, the content of patent material is maintained and stored in full-text and XML formats. However, a formal and unambiguous semantic representation would be essential for both, human and machines, to facilitate retrieval, interpretation, and analysis of patent material and, on the other hand, make the examination and classification tasks for the specialists at the patent offices much more straightforward. There is strong evidence that in the long run the availability of an appropriate semantic content representation will culminate in the compilation of patent knowledge bases.

In this paper, we present a first approach to the representation of patent material based on an ontology framework. This framework attempts to address the limitations of the current formats in that it accounts for a semantic representation as well as for other knowledge aspects of patent material, such as inferring the status of a patent (granted, lapsed, withdrawn, etc.) based on the legal status events published by the offices. Thus, it also naturally provides a homogeneous representation of patent information merged

from different sources and services. The framework is the result of the still ongoing work within the PATExpert project [7]. The overall scientific objective of PATExpert is to design and implement a patent content representation formalism based on Semantic Web technologies for selected technology areas and to develop techniques that are grounded in this formalism: retrieval, classification, extraction, summarization, visualization, and assessment of patent material.

The rest of the paper is structured as follows. Section 2 gives an overview of various types and characteristics of patent information and presents the overall ontology design in PATExpert. Sections 3–5 describe the ontologies realized in the framework, and Section 6, finally, draws the conclusions and provides an outlook to future work.

2. Overview of the Ontology Framework

Patent documentation constitutes a diverse source of information for different groups of users that have to fulfill different tasks. For example, patent examiners have to investigate patent applications for its novelty, scientists have to search for the state-of-the-art in a specific technological area, or companies have to proof the validity of newly granted patents of their competitors. To represent the different aspects of patent information in terms of a formal and unambiguous semantic representation means to interpret patent documents as *knowledge objects*. The definition of knowledge objects in terms of concepts and relations between them is separated from their instantiation. This definition is realized using ontologies, whereas the knowledge instantiation is managed by means of a knowledge base. In PATExpert, the Web Ontology Language (OWL) [13] is used as formalism to encode the ontologies. The knowledge base is realized as an RDF store [5] in tandem with a relational database backend. In this paper, we focus, first of all, on ontologies.

A closer look at the representative patent material reveals that the knowledge on patent documentation can be divided into three major blocks: (i) common sense knowledge, e.g. definitions like a *Patent IS-A Certificate* that confers a right or obligation on the holder of the Certificate, (ii) patent genre-specific knowledge, and (iii) domain-specific knowledge. In order to address these different aspects, our ontology framework consists of several modules. For the design of the ontology framework, current patent standards and rules, such as the WIPO standards [11] and the *European Patent Handbook*, as well as best-practices for ontology design [18,16,9] have been taken into consideration. Fig. 1 gives an overview of the ontology modules. A more general overview of the PATExpert representation formalism is given in [19].

To capture common sense knowledge in patents, we use the *Suggested Upper Merged Ontology* (SUMO) [17], to which all other modules are linked. One of the reasons why SUMO has been chosen as the upper level ontology is that it is linked to the English lexical ontology WordNet [15]. WordNets for the European patent languages English, French and German form our linguistic ontologies. Patent genre specific knowledge is encoded by means of the *Patent Upper Level Ontology* (PULO), which subsumes the *Patent Metadata Ontology*, the *Patent Structure Ontology*, and the *Patent Drawings Ontology*. The *Domain Ontology* comprises the so-called *Core Domain Ontology*, the *Auto Domain Ontology* and the *Patent Classification Ontology*.

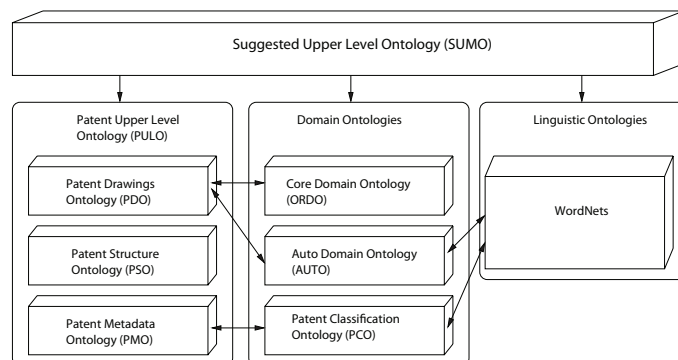


Figure 1. PATEXpert Ontology Modules

3. Patent Genre Specific Ontologies

3.1. Patent Metadata Ontology

Patent metadata further describe patent documents or related data; they can be explicit or implicit. Both explicit and implicit metadata are crucial for patent search. *Explicit metadata* include bibliographic data such as the title, the name of the applicant, the date of filing and publication, etc. *Implicit metadata* (such as, patent or literature citations within a patent, kind of patent (e.g., a process patent or a product patent), etc.) must be extracted from higher level associations between patent documents or from their textual content. Major metadata aspects modelled by the Patent Metadata Ontology (PMO) are bibliographic data, patent family information, legal status information, and concepts for generic semantic or textual annotations (fig. 2 metadata).

Bibliographic data are defined in the World International Property Organization (WIPO) Standards [11], in particular ST.9, ST.32, and ST.36, which can be adapted to national and regional patent laws and conventions. ST.9 defines about 60 data entities grouped in eight major groups that describe for example the identification of a patent, data relating to priority under the Paris Convention, date of making the patent applications available to the public, references to other legally or procedurally related patent documents, etc. In PMO, most of this information is modelled in terms of object and data properties.

A *patent family* encompasses all patents that belong to the same invention.¹ Various definitions of a patent family exist. The narrowest definition (as, e.g., in *esp@cenet*) considers a family to include only those documents whose priorities and claims match.² In PMO, patent families are modelled using the *sameFamily* property. The transitive closure of patent documents related to other documents via *sameFamily* builds up a patent family. Since there are different definitions of patent families, each specific definition is modelled as a sub-property of *sameFamily*.

¹For different reasons (e.g., because of the protection in different countries, division or continuation of an application, etc.), one invention may have multiple documents.

²The first filing of a patent application may be considered the priority application and referred to from other applications.

Legal status information describes all significant steps in the lifetime of an invention. It includes data such as change of owner, examination request, grant, revocation etc. As its life cycle proceeds, a patent application usually goes through different legal stages. In PMO, legal status information is modelled using a simple event model. Each *Event* instance can have an associated set of *AttributeValue* instances. An event can trigger a *Transition* which has a resulting *State* and an optional output, which is an instance of *MultimodalDocument*. For instance, an examination request event triggers an examination, which results in a new *examination in progress* status and eventually has an output *examination report*.

3.2. Patent Structure Ontology

The Patent Structure Ontology (PSO) aims to capture the structural decomposition of a patent document. Every patent has clearly defined units that are determined by patent standards. A patent application has to encompass parts such as description, claims, drawings and abstract. Furthermore, each part has its own structure which needs to be taken into account. Thus, headings, titles, paragraphs, subsections are important for patent document structure analysis.

PSO also models the multimodal aspect of patent content. Structurally, a patent contains the textual description of the invention and the figures. Figures may be diagrams, flow charts, waveforms or technical drawings. Additionally, one can find other embedded objects such as tables or formulae. All these multimodal elements may contain characters, words, numbers or special symbols. There can be references between text and figures most notably coming from the numbering of captions, which provide correspondence to the diagrams, and from the numbers appearing in the textual descriptions, which correspond to the numbered components of the drawings (section 4.5 gives a detailed example).

Fig. 2 (patent structure) gives an overview of the major concepts in PSO and shows the relations between higher-level SUMO and PSO concepts. The central concept is the concept of *PatentApplication*, which can be augmented by *PatentSupplement* or *Non-PatentSupplement* instances. An application can be made public which results in a *Patent-Publication*. Furthermore, an application is associated with *PatentContent* instances that model the different content types of an application, e.g. *Description*, *Claims*, *Drawing*, etc. Each *PatentContent* instance can have citation references, that are modelled using a *Citation* relation that allows to define an optional relevance indicator for the citation, e.g. for citations in a search report. Each content part can also define internal references modelled as *Reference* relations. There are specialized sub-properties of *hasContent*, such as *hasDescription*, *hasClaims*, *hasDrawing*, etc., which are omitted in Fig. 2. For better readability specific sub-classes of *PatentApplication* are also omitted.

3.3. Patent Drawings Ontology

Drawings constitute an essential component of patent documentation. The goal of the Patent Drawings Ontology (PDO) is two-fold: (i) to provide a taxonomy for patent drawings types, (ii) to enable the explicit expression of the drawings content. There are several types of patent drawings: system diagrams, circuits, waveform, flowchart, etc. PDO defines a concrete hierarchy of patent drawings classes that can be exploited in order to improve search precision and recall performance in patent drawing retrieval tasks. Fig. 2

(patent drawings) displays the drawings taxonomy proposed by the PDO. Furthermore, the association of patent drawings with concepts from the domain ontology is possible via the *conveysInformationSubject* relation. In that way, it is possible to express the content and type of the object that is depicted in the drawing.

4. Domain-Specific Ontologies

Our work on domain-specific ontologies is divided into two parts: the construction of the Patent Classification Ontology (PCO), which is based on the International Patent Classification (IPC) [4], and the construction of ontologies for two technical areas on which we evaluate our knowledge-intensive approach to patent processing: *Optical Recording Devices* and *Machine Tools*. In this paper, we elaborate on the methodology followed to build the *Optical Recording Domain Ontology* (ORDO). The ontology was engineered in two phases. First, a relatively small ontology (the *Core Domain Ontology*) was manually built, then an extension of the manual ontology (*Auto Domain Ontology*) was acquired by automatic ontology learning. Obviously, the ultimate goal of this effort is to populate the resulting ontology with instances of concepts describing the actual content of the patents.

4.1. Patent Classification Ontology

Patent classifications serve as an instrument for the orderly arrangement of patent documents in order to facilitate access to the technological and legal information contained therein. Classifications are also a basis for selective dissemination of information to all users of patent information for investigating the state of the art in given fields of technology and for the preparation of industrial property statistics which in turn permit the assessment of technological development in various areas.

The most prominent patent classification is the International Patent Classification (IPC), which has been developed by the World Intellectual Property Organization (WIPO) for more than 20 years. It is used by almost all patent offices for the classification of patents. The IPC has mappings to other classification schemes, such as the European Patent Classification (ECLA) [1], the US Patent Classification (USPC) [10] or the Japanese FI/F-Term classification [3].

The *Patent Classification Ontology* (PCO) allows general mappings between concepts of classification schemes (such as IPC, ECLA, etc.) and patent instances. The PCO further allows the integration of user defined classifications, which is a very important aspect, e.g. for highly specialized companies that on the one hand need a very detailed classification but on the other hand only work on a very limited subset of the IPC. Also adhoc annotation of patent documents can be seen as kind of 'light-weight' classification, that helps users to group patents under certain aspects.

The PCO is based on the meta classification concepts defined in PULO. Fig. 2 (Classification) shows the most important meta concepts for PCO and an adaptation to the IPC schema. In PCO classification mappings (between different classification schemes) can be modelled as well as relationships between concepts within on classification scheme. In the first version, the PCO has been generated using the IPC Version 8 from 01/01/2007. It comprises 69544 classes, among them 13,095 that have references to other PCO classes and 56,449 classes without references.

4.2. Core Ontology: The manual construction of ORDO

The initial manual domain ontology has been built in two phases. In the first phase, we exploited the part of the IPC dealing with Optical Recording. Given that the IPC has been devised and is actively maintained by domain experts, it is an excellent source of knowledge. The knowledge is implicit in the structure of the classification, and can be easily recognized and reconstructed in an ontology. Consider, e.g., the following fragment of IPC (representing a category and its sub-category).

G11-B7/002:	Recording, reproducing or erasing systems characterised by the shape [N: or form] of the carrier
G11-B7/0025:	with cylinders or cylinder-like carriers [N: or cylindrical sections or flat carriers loaded onto a cylindrical surface], e.g. truncated cones

From this description, one can induce that shape is an important feature of a record carrier (this feature is indeed used to organize the sub-categories of G11B7/002), and that “cylindrical” is one possible shape of a record carrier. The elicitation process, i.e., the process of inducing the knowledge, has been performed by manually encoding the knowledge implicit in the classification into ontological axioms. Thus, from the labels of the category G11B7/002 and its sub-category G11B7/0025, we infer the following axioms:

$$\text{record_carrier} \sqsubseteq \exists \text{has_carrier_shape}.\text{record_carrier_shape}$$
$$\text{record_carrier_shape}(\text{cylindrical_shape})$$

The first axiom formalizes the fact that a record carrier has a shape; the second axiom formalizes the fact that cylindrical is one of the possible shapes of a record carrier. Note that from the classification it is not clear that such a shape is unique. i.e., that a record carrier has a unique shape, so the axiom states only that there is one shape and does not exclude the situation in which a record carrier has more than one shape. Such an elicitation is done manually by knowledge engineers without the intervention of domain experts, relying only on commonsense knowledge.

In the second phase, a corpus of about 5.000 patents classified under the class G11B7 has been used.³ By using simple statistical techniques, an ordered list of the most frequent terms in the corpus was build (including both mono- and bi-grams). The top-most part of this list was scrutinized to check the coverage of the IPC-based domain ontology, and new (i.e., not previously defined) concepts were added. The definition of concepts not included in the IPC was based on glossaries and Web Sites such as <http://www.answers.com/>.

Currently, the manual portion of ORDO contains 210 classes and 57 object properties. Domain classes are for example `head` (the class of recording heads) or `flying-type_head` (the class of recording heads which are of type flying). Object properties are for example `supports_recording_method` (the relation between

³G11B7 is specified as follows: **Recording or reproducing by optical means**, e.g. recording using a thermal beam of optical radiation [N: by modifying optical properties or the physical structure], reproducing using an optical beam at lower power [N: by sensing optical properties]; Record carriers therefore; ...

a tool and the supported recording method), `has_carrier_shape` (the shape of a record carrier), or `detection_target` (the relation between a detection action and the detected object).

4.3. *Auto Domain Ontology: Ontology Learning*

ORDO does not cover the big number of concepts and properties that are necessary to represent the content of a patent (or at least the most important part of it). Given the PATExpert application scenario, it is totally unrealistic to define such a big number of concepts manually. For this reason, we had to resort to automatic ontology learning. To illustrate the technical details of the specific ontology learning techniques that we implemented is outside the scope of the present paper. Let us only mention that we followed two main approaches. The first is based on projecting relevant sections (is-a paths) of the WordNet hierarchy into ORDO. The second is based on searching the Internet for textual patterns expressing is-a relations. Automatically acquired concepts are asserted in a separate ontology (the *Auto Domain Ontology*, AUTO), which is linked but distinct from ORDO and which currently includes around 5.000 concepts.

4.4. *Ontology Population*

The ORDO and AUTO ontologies are populated as a result of a relation extraction algorithm, exploiting, for English, the output of the MiniPar dependency parser. Again it would be unfeasible to extract all the possible relations contained in a patent, so we focused on a restricted number of them, with a special emphasis on relations expressing meronymic and movement relations between components. It also turned out very useful to extract relations related to image descriptions which are very common in the text of patents. Such relations can be used to infer the type and the content of the images which form an essential part of any patent.

4.5. *Patent Knowledge Storage and Retrieval*

All instances pertaining to a particular patent document are considered to belong to a single RDF graph. In addition, the instance names (subjects and objects) are constructed in such a way that the position of the text referring to them can be derived, e.g., the name `2_2-2_3` denotes that the second and third tokens of the second text sentence, refer to the particular instance. Such information is available from and utilized by the linguistic analysis tools employed for the patent text processing. The knowledge encoding approach proposed by the authors is illustrated by means of the following example (figure 3).

The text appearing below the figure has been extracted from the patent by linguistic analysis techniques. A part of the knowledge which is conveyed by the figure and the text associated with it is encoded in the following RDF-triples:

```
_:US3528287:25_2-25_3 rdf:type pulo:Figure .
_:US3528287:25_2-25_3 pulo:conveysInformationSubject _:US3528287:25_5-25_6 .
_:US3528287:25_2-25_3 pulo:conveysInformationSubject _:US3528287:25_18-25_19 .
_:US3528287:25_5-25_6 rdf:type ordo:disk .
_:US3528287:25_5-25_6 pulo:hasFigureElementReferenceLabel "12"xsd:string .
_:US3528287:25_5-25_6 ordo:hasComponentMaterial _:US3528287:25_9-25_9 .
_:US3528287:25_9-25_9 rdf:type auto:Plexiglass .
_:US3528287:25_5-25_6 ordo:rotatedBy _:US3528287:25_18-25_19 .
```

_:US3528287:25_18-25_19 rdf:type ordo:Motor .

_:US3528287:25_18-25_19 pulo:hasFigureElementReferenceLabel "14"xsd:string .

The above RDF triples encode the fact that the tokens "FIG. 3" of the text refer to an object of type Figure (defined in PULO) and that the figure conveys information about two objects. The first of those two objects "disk 12" is of type Disk (defined in ORDO) and has label "12". It is also made of the material "Plexiglass" (defined in AUTO) and is rotated by a second object. This object "motor 14" is of type Motor (defined in ORDO) and has label "14".

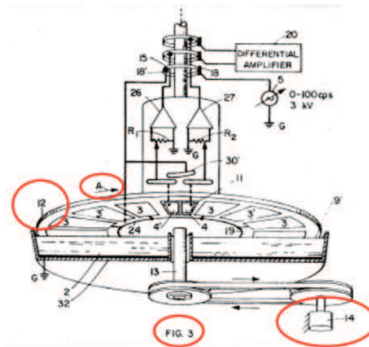


Figure 3. Figure from US3528287 referenced as 'In FIG. 3 the disk 12 may be plexiglass or some other dielectric material rotated by the motor 14 in the direction of the arrow A.'

5. Linguistic Ontologies

The term "linguistic ontology" is used to refer to two quite distinct kind of objects. Firstly, a linguistic ontology may be a formal representation of linguistic meta-level objects (henceforth, "meta-level linguistic ontology"); cf. the *Generalized Upper Model* (GUP[14]) for an example. Secondly, a linguistic ontology is a formal representation of lexical information (henceforth, "lexical linguistic ontology"). In this case, the concepts of the ontology represent meanings of words used in some language. WordNet [12] is the typical example of a lexical linguistic ontology. Lexical linguistic ontologies are essential for knowledge-oriented text processing approaches such as the one adopted in PATExpert, which aims at producing conceptual representation of patents. WordNet-like ontologies provide the bridge between words that appear in texts and concepts that are the building blocks of such conceptual representations. We exploit existing WordNets for English, German, Spanish, and French. Given that WordNets usually cover general purpose usage of language and patents contain many domain specific terms, we need to add at least a part of such technical terms to the existing hierarchies, by exploiting ontology learning techniques.

Due to the number of concepts included in word-level linguistic ontologies, their exploitation poses a number of practical questions. Thus, the English WordNet contains more than 150.000 lexical concepts (called "synsets"). Although an OWL version of WordNet is available [8], this is not the most efficient representation of it. For this reason, we are not including WordNets explicitly in the PATExpert ontology. Instead, WordNet information is maintained in a database, and pointers are created from the lexical concepts in the data base and the relevant concepts in the PATExpert ontology. Most of the links concern domain ontology concepts.

6. Conclusions and Future Work

This paper presented a complex ontological framework for the representation of patent documentation. The necessity for developing such a framework stems from the heterogeneity of patent material that is available through state-of-the-art patent services, the multimodal and multilingual nature of it, as well as the need to express semantics in order to enhance patent retrieval and analysis tasks. The proposed framework provides the necessary infrastructure for the next-generation patent retrieval and analysis services. There are already components, developed within the PATExpert project [7], that are based on the proposed framework. The ongoing efforts for the deployment of the framework in a real-world situation are faced with the challenge of scaling the patent retrieval and analysis system to cope with the current magnitudes of available patent documentation. Careful engineering of the RDF storage, querying and inferencing mechanisms will be required to integrate the proposed infrastructure with the workflows currently established in patent offices and companies worldwide.

References

- [1] ECLA - European Classification. <http://v3.espacenet.com/eclasrch>.
- [2] esp@cenet homepage. <http://ep.espacenet.com>.
- [3] FI/F-term Search. <http://www4.ipdl.inpit.go.jp/Tokujitu/tjftermena.ipdl?N0000=114>.
- [4] International Patent Classification 8th Ed. Vol. 5 Guide. http://www.wipo.int/classifications/ipc/en/other/guide/guide_ipc8.pdf.
- [5] Jena - A Semantic Web Framework for Java. <http://jena.sourceforge.net/>.
- [6] Open patent services homepage. <http://ops.espacenet.com/>.
- [7] Patexpert project website. <http://www.patexpert.org/>.
- [8] RDF-WORDNET. <http://www.w3.org/TR/wordnet-rdf/>.
- [9] Semantic Web Best Practices and Deployment Working Group. <http://www.w3.org/2001/sw/BestPractices/>.
- [10] USPC - United States Patent Classification. <http://www.uspto.gov/go/classification/>.
- [11] Wipo standards. <http://www.wipo.int/scit/en/standards/standards.htm>.
- [12] WORDNET. <http://wordnet.princeton.edu/>.
- [13] OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004, 2004. <http://www.w3.org/TR/owl-guide/>.
- [14] J. Bateman, R. Henschel, and F. Rinaldi. Generalized upper model. Technical report, GMD/Institut für Integrierte Publications- und Informationssysteme, 1995.
- [15] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge (MA): MIT Press, 1998.
- [16] Aldo Gangemi. Ontology design patterns for semantic web content. In Y. Gil et al., editor, *ISWC 2005*, pages 262–276. Springer, 2005.
- [17] I. Niles and A. Pease. Towards a standard upper ontology. In *2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, 2001.
- [18] J. F. Sowa. *Knowledge Representation*. Brooks/Cole, 2000.
- [19] L. Wanner, S. Brüggemann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki. Towards Content-Oriented Patent Document Processing. *World Patent Information*, 2007 (to appear). doi:10.1016/j.wpi.2007.03.008.