Application of Semantic Technologies for Representing Patent Metadata

Mark Giereth, Achim Stäbler, Sören Brügmann*, Martin Rotard, Thomas Ertl
Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart
Universitätsstr. 16, 70569 Stuttgart, Germany
{giereth|staebler|rotard|ertl}@vis.uni-stuttgart.de
*Industrie Software Jochen Brügmann
Bokeler Str. 18, 26871 Papenburg, Germany
soeren.bruegmann@bruegmann-software.de

Abstract: Patents belong to the few types of public information that have a big impact on national and international economies. During the last years there have been great efforts in making patent data available electronically for the public via online services. But today's services provide heterogeneous data structures which makes automatic processing difficult. None of the services supports all user aspects, so that different services have to be combined. In this paper we present an ontology-based approach for representing patent metadata and describe a *Patent Metadata Ontology (PMO)* that models the major aspects of patent metadata. The advantage of our approach is to provide a homogeneous representation of patent metadata merged from different sources. It allows for identifying context and dependency information more easily than today's database-centric structures and interfaces.

1 Introduction

Patents are of great importance for national and international economies. But they are also a valuable source of up-to-date scientific and technological information. Patent applications have to be published after a defined time period. This ensures that their content is publicly documented. It is widely assumed that the worldwide stock of patents comprises a large part of all scientific and technical knowledge. The total number of patent documents worldwide is estimated to be about 60 millions[1].

The negative consequence of the rapidly growing number of patents is an increasing opaqueness of the patent market that makes it difficult for smaller companies living from their inventions to succeed in the market. The risk of patent litigation increases which is reflected by a growing number of press articles on law suits related to patent right violations. Given that the amounts at stake in such disputes often surpass several million euros, patent litigation is a serious threat to the existence of numerous companies. During the last years there have been great efforts in making patent data available electronically for the public. But finding relevant patents, e.g. related patents of competitors, is still a

[1] Source: http://ep.espacenet.com/ep/en/helpV3/espacenet.html

complex task. It needs experts that are familiar with the specific patent and domain terminologies. The search is usually done by combining specific terms that occur in the title or the abstract of a patent with metadata details, such as name of the applicant, publication date, citations, priorities, legal status, or classification information. Patent metadata also play an important role for patent valuing, for tracing competitors and for identifying their strategies. Today, patent metadata are made available by various online services. The esp@cenet and epoline services [ESP, EPO] offered by the European Patent Office (EPO) are two prominent examples of such initiatives. But the services provide different data sets and even use different data structures, which results in a heterogeneous view on patent metadata. This makes it difficult for automatic processing.

This paper describes a new ontology-based approach of representing patent metadata using the Resource Description Framework (RDF) and the Web Ontology Language (OWL). By introducing the *Patent Metadata Ontology (PMO)*, our overall goal is to provide a semantically well-defined and homogeneous representation for the major kinds of patent metadata. The advantage of our approach is to preserve the context information rather than providing database-centric structures and interfaces. Also the merging of different pieces of data from different services is facilitated by applying the well-established framework provided by RDF.

There are various studies on patent document and metadata analysis (cf. [JT02, Mar02]). But to our knowledge, there have been no studies on ontology-based representation of patent metadata. An ontology for the more general intellectual property rights topic has been discussed in the context of digital rights management [Del02]. Unfortunately, the aspect of patent metadata has not been addressed by that ontology.

The next section gives an overview of the major kinds of patent metadata. Section 3 briefly describes the design of the Patent Metadata Ontology (PMO). In section 4 the population of PMO and the mapping to XML patent documents is described. The paper ends with conclusions and future work.

2 Patent Metadata

Data that describe or are related to patent documents are called patent metadata. We can distinguish between explicit and implicit metadata. *Explicit metadata* is given in the front page of a patent document and include bibliographic information like title of the invention, inventor name, classification, countries in which the invention is to be protected, etc. *Implicit metadata* has to be extracted from higher level associations between patent documents as well as from their textual content, for example patent or literature citations occurring in the patent content or the patent type extracted from the claims (e.g. a process patent or a product patent).

Patent metadata can further be classified into internal and external data. *Internal metadata* can be derived from a single patent document, whereas for *external metadata* other patent documents or data sources have to be taken into consideration. Examples for external metadata are events concerning the legal status of a patent or additional applicant

or inventor information. In the following, we briefly describe three major areas of patent metadata: bibliographic data, patent families, and legal status information.

Bibliographic Data

Bibliographic data for various kinds of patent documents are described by the World International Property Organization (WIPO) standards ST.9 (Recommendation Concerning Bibliographic Data on and Relating to Patents and Supplementary Protection Certificates (SPCs)), ST.32 (Recommendation for the Markup of Patent Documents using SGML) and ST.36 (Recommendation for the Processing of Patent Information using XML) (for a list of all WIPO standards see [WIP]). Bibliographic data are further defined by national or European patent laws and conventions. WIPO ST.9 defines about 60 data entities widely used on the first page of patent documents or in patent gazettes. Each metadata entity is associated with a unique two-digit INID code (Internationally agreed Numbers for the Identification of bibliographic Data) describing eight major groups:

- Identification of the patent, SPC or patent document (1x)
- Data concerning the application for a patent or SPC (2x)
- Data relating to priority under the Paris Convention (3x)
- Date(s) of making available to the public (4x)
- Technical information (5x)
- References to other legally or procedurally related domestic or previously domestic patent documents (6x)
- Identification of parties concerned with the patent or SPC (7x)
- Identification of data related to International Conventions other than the Paris Convention, and to legislation with respect to SPCs (8x and 9x)

Patent Families

The first filing of a patent application in some patent office is considered the priority application. Priorities form a special kind of relationships between patents and are of great interest in patent analysis. The most recognized concept in this area is the concept of patent families. A patent family encompasses all patents belonging to the same invention. For different reasons one invention can be protected by multiple patents. The main reason is that a patent is only valid for one country. Thus, protection for different countries results in multiple patents describing the same invention. Further an applicant can be forced by the patent office to divide his application, if it describes more than one invention. This is because some patent laws provide that one patent must disclose only one invention.

There are various definitions on what constitutes a patent family. The narrowest definition of a patent family is the definition that considers a family to include only those documents

whose priorities and claims match. This definition is used by esp@cenet . A broader definition involves those cases, where the applications have at least one priority in common.

Legal Status Information

The legal status describes all significant steps in the lifetime of an invention, from first publication (in some cases even from the filing) to the end of term of the patent, and includes data such as change of owner, examination request, grant, revocation etc.

The EPO keeps a history of these so called legal status events pertaining to a patent in the INPADOC (International Patent Documentation Center) database which contains more than 32 million legal status data records from 46 patent issuing organizations since 1978. This information is searchable via the esp@cenet and epoline services. A list of currently about 3.000 internationalized legal status events is published at the INPADOC website[2]. The list is regularly updated on a weekly base. A legal status event consists of a name, a date, an event code, an optional country code and optional attribute-value pairs.

Metadata Services

This section lists a set of online services provided by the EPO for retrieving patent documents and patent metadata. The list is not intended to be exhaustive.

- esp@cenet [ESP] is aimed at patent information end-users. It contains all the patent documentation available to EPO examiners and the latest patent applications from all the EPO member states.
- epoline [EPO] is the name given to the range of online products and services designed by the EPO to allow patent applicants, attorneys and other users to conduct their business with the EPO electronically. The epoline includes Register Plus – a service that provides legal status, event history, citations, patent family and application documents. Search results can be selectively downloaded in XML format.
- Open Patent Services (OPS) [OPS] provide a web service based interface to bibliographic, patent family and legal status data.
- European Publication Server [EPS] provides online access to the collection of European patent documents published by the EPO (in XML-ST.36 since 2006).

These services define their own XML data structures which are related to ST.36 in some kind. Each service provides one or more aspects of the metadata description of a patent. None of the services provides all aspects, so that as a result different services have to be combined (see section 4).

[2] `http://www.european-patent-office.org/inpadoc/stats/`

3 Patent Metadata Ontology

In this section we describe the design of the Patent Metadata Ontology (PMO). PMO is one out of a set of ontologies developed within the PATExpert (Advanced Patent Document Processing Techniques) project [Pat] focusing on the bibliographic data, patent families, legal status information, classificational information, citations and generic annotations. Figure 1 gives a brief overview of the most relevant concepts for describing these kinds of information. For better readability, datatype properties and property names are omitted.

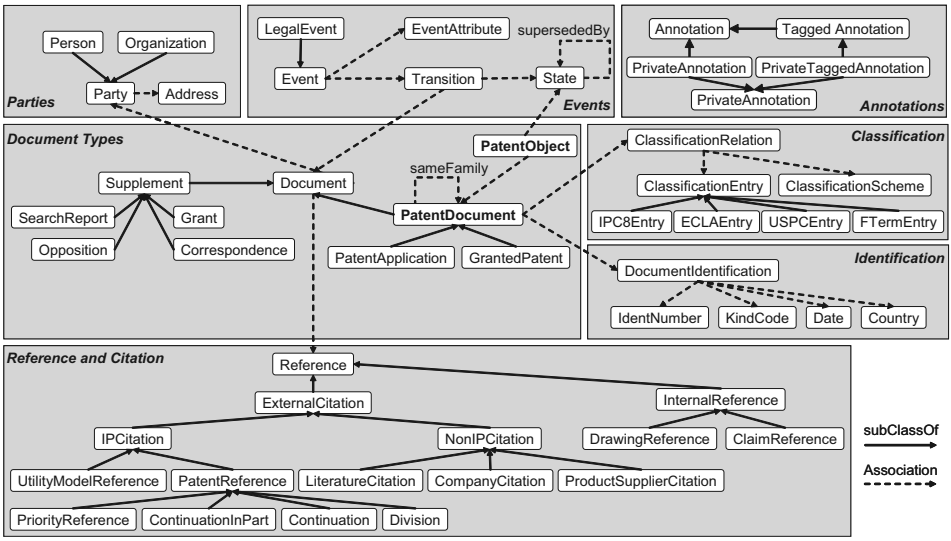

Figure 1: Overview of Patent Metadata Ontology concepts

Bibliographic Data

Bibliographic data mainly describe relations between patent documents and dates, parties, states, documents, etc. In PMO, an abstract `PatentObject` is described by one or more `PatentDocument` instances. Each `PatentDocument` is uniquely identified by a `DocumentIdentification`. A published patent application for example is identified by the four components: country code (according to WIPO ST.3), document number (according to WIPO ST.6), kind of document code (according to WIPO ST.16) and publication date. A patent document has associated parties like applicants, inventors, attorneys, etc. which can be persons or organizations like companies.

As a subclass of `Document`, a patent document can have references to other other documents or document parts. The reference can cite other documents, e.g. literature citations, or can refer to other parts in the same document, e.g. references to drawings or references to claims. `InternalReference` instances need to know the document structure, which is modeled in detail in the Patent Structure Ontology, another ontology developed within

PATExpert. Further, each patent document can be supplemented by other documents like search reports, grant certificates, oppositions, etc. depending on the type and the state of the document. This is modeled by associating an appropriate sub-class of `Supplement` to the patent document.

Classifications

Each patent document is classified using one or more classification schemes like the International Patent Classification (IPC), the European Patent Classification (ECLA) or the US Patent Classification (USPC). Most of the patents are classified using the IPC. The IPC is a hierarchical system in which all technology areas are organized into a range of sections, classes, subclasses, groups and subgroups. The current version of the IPC, version 8, has been divided into a Core Level with about 20.000 classes and an Advanced Level with about 60.000 classes. While the core level is intended to remain stable, the advanced level is going to be revised every 3 months.

In PMO, each patent document is classified by one or more `ClassificationRelation` instances. This n-ary relation defines the classification and the used classification scheme. Depending on the underlying scheme, an appropriate `ClassificationEntry` sub-class, e.g. `ECLAEntry` for the European Patent Classification, is used to describe the classification.

Patent Families

Patent families are described by the generic `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 modeled as a sub-property of `sameFamily`. INPADOC for example defines a family as a set of documents having at least on priority in common, whereas esp@cenet defines a family as a set of documents where all priorities are the same. In PMO, INPADOC families are modeled using `sameINPADOCFamily` and Esp@cenet families by using `sameEspacenetFamily` properties.

Legal Status

Events and in particular legal status events are modeled using a simple event model as presented in figure 1. Each `Event` instance can have an associated set of `AttributeValue` instances[3]. An event can trigger a `Transition` which has a resulting `State` and an optional outcome. The outcome of a transition is a `Document` instance. For example, an *examination request* event triggers an *examination* which results in a new *examination in progress* status and eventually has an output *examination report*.

[3] A more elaborated event taxonomy is planned in future versions.

4 Ontology Population and Mapping

For populating PMO, we have started with 8.000 European patents in the domain of mechanical engineering. The patent metadata have been downloaded using the Open Patent Services for the family and legal status information and the Register Plus service for the bibliographic data. Both services use document type definitions that differ from the WIPO ST.36 standard. For a subset of the patents we got the ASCII full-text from the European Patent Office. For patents filed since 01/01/2006 we retrieved the fulltext in ST.36 conforming XML format using the European Publication Server.

The ontology population process has two phases. The first phase includes the generation of RDF/XML from the Open Patent Services and Register Plus XML documents. This is done by using XSLT stylesheets. In the second phase an RDF store is filled using the results of phase one. The major task in this phase is to associate the different patent documents to the corresponding abstract patent object. Since a patent object has no unique identifier, it is represented as an RDF blank node. Whereas patent document instances are identified by an URN generated from the patent identification information. For a published patent application for example, the URN is composed as follows:

```
urn:pat:<Country-Code>-<Number>-<Kind-Code>-<Date>
Example: urn:pat:EP-0581199-B1-20060524
```

In order keep the knowledgebase small, we follow a hybrid approach. The document structure skeleton is represented in RDF, whereas the content is stored separately and linked to the metadata ontology by making extensive use of the XPointers. For that, we provide a mediator service that returns the patent content based on a given URN (as described above). Within the XML ST.36 conforming documents, each information item is identified by a unique ID attribute. So for example linking to a claim item with the ID '002' can simply be realized by using the XPointer:

```
urn:pat:EP-0581199-B1-20060524#xpointer(id('002'))
```

Using the described approach has showed, that a flexible population of a patent metadata knowledge base with data retrieved from different sources can be achieved by combining semantic technologies with XML technologies.

5 Conclusions and Future Work

We have presented a new ontology-based approach for representing patent metadata. The advantage of our approach is to provide a homogeneous representation of data that is currently represented by various XML schemes and services. The ontological approach together with additional reasoning capabilities will allow for flexible and extensible patent applications, e.g. for valuing, searching and visualizing of patent material. RDF query languages will allow for freely combining and filtering patent metadata and thus to build up user defined views and analyses. Important aspects in this regard are in particular patent citations, inventor data and classifications [JT02]. The study of citations is supposed to be

valuable to reveal technology development. Inventor data can be used to identify links between patents through co-authorship. Also address information can to be useful, for example to map knowledge flows geographically.

PMO is one of a set of ontologies developed within the PATExpert project [Pat]. It is directly linked to a Patent Structure Ontology and a Patent Upper Level Ontology. An important objective of PMO is to improve the integration of patent metadata with other PATExpert ontologies in order to develop a uniform representation formalism for patent material.

An essential goal for future work, will be the development of techniques for visualization of and navigation in large patent knowledge networks. Further, the development of methods for annotating patent documents and metadata will be another objective that is going to be addressed in the future.

Acknowledgements

Research reported in this paper has been partly funded by the European Commission within the PATExpert (Advanced Patent Document Processing Techniques) project. PATExpert's overall scientific objective is to develop a patent content representation formalism based on Semantic Web technologies for the selected technology areas in optical recording and mechanical engineering and to investigate the retrieval, classification, multilingual generation of concise patent information, assessment and visualization of patent material encoded in this formalism.

References

[Del02] J. Delgado et. al. An ontology for Intellectual Property Rights: IPROnto. In *Posters of the ISWC*, 2002. http://iswc2002.semanticweb.org/posters/delgado_a4.pdf.

[EPO] epoline. http://www.epoline.org.

[EPS] European Publication Server. http://patentinfo.european-patent-office.org/off_pubs/pub_serv/.

[ESP] esp@cenet. http://ep.espacenet.com.

[JT02] A. Jaffe and M. Trajtenberg. *Patents, Citations and Innvocations*. MIT, 2002.

[Mar02] K. Markellos et. al. Knowledge discovery in patent databases. In *Proceedings of the CIKM '02*. ACM Press, 2002. http://doi.acm.org/10.1145/584792.584915.

[OPS] Open Patent Services. http://ops.espacenet.com/.

[Pat] PATExpert Project Website. http://www.patexpert.org/.

[WIP] WIPO Standards. http://www.wipo.int/scit/en/standards/standards.htm.