

Visualization Enhanced Semantic Wikis for Patent Information

Mark Giereth and Thomas Ertl

Visualization and Interactive Systems Institute, University of Stuttgart
{giereth|ertl}@vis.uni-stuttgart.de

Abstract

In this paper we present a new approach for using semantic wikis for collaborative patent search and annotation. We describe an extension that allows integrating interactive visualizations into semantic wikis for getting deeper insights into the classificatory, geographical, and temporal distribution of large patent sets. This approach differs from typical wiki usage scenarios in the sense that it combines automatic content generation based on patent search activities of the users with user driven semantic annotation of patent information, e.g. patent rating, linking with prior art, reviews, translations, discussions, etc. The content generation involves a semantic model that is described in terms of different ontologies for patent information. A modern wiki system is used for semantic annotation, comments, discussions, versioning, notification, and full-text search. Our approach is motivated by using available functionalities of a modern wiki system in combination with visualization techniques to directly implement major user requirements for supporting the knowledge-intensive tasks of patent search and understanding.

1. Introduction

The original intent of the worldwide patent system is to encourage research and development (R&D) and to promote the disclosure of innovations, and thus, to avoid unnecessary R&D costs by making the latest technology information publicly available. Therefore, for a working patent system the aspect of public availability of patent data is crucial. During the last years there have been great efforts by the major patent offices in making the patent full-texts, images, and metadata publicly available.

In this paper we describe a system for visualization enhanced semantic wikis that is used as user interface to patent information. The system is part of the prototype developed within the PATExpert [1] project. The overall approach for the development of

the system described in this paper is characterized by the following aspects: First, to use a semantic model for the representation of patent information and user annotations. Second, to use semantic wiki technology for adding semantic annotations to the original patent data within a community process. And third, integration of interactive visualizations for classificatory, temporal, geographic, and other patent information aspects

In traditional wikis content pages are created, edited and annotated by users. In our system content pages are automatically generated based on the patent search activities of users. Additionally to its textual representation, semantic concepts and metadata are integrated using Microformats [2] and RDF (Resource Description Framework) [3]. Generated pages cannot be edited directly, because they contain immutable information published by a patent office. However, commenting and semantic annotation of pages is supported, e.g. discussions of patent content, patent reviews, patent rating, linking to prior art, etc.

Our approach is motivated by using the functionality of modern wiki systems to directly implement major user requirements: the possibility to annotate content, to manage different versions (how do the query results change over time), and to send notifications when content has been updated. The proposed system combines user interactions (search and annotation) with automatic actions (adding new patents to the system, repeating searches on a weekly base, notifying users, etc.).

2. Related Work

Patent information systems are basically bibliographic systems specialized for information associated with patents. The most prominent commercial patent databases are Derwent World Patents Index, STN, and PatBase. They come with their own search tools and viewers. Commercial databases provide ‘added values’ by integrating data from different patent offices and by providing abstracts and translations of new patents. Users have to

pay if they want to access the data or use the tools. Commercial patent data providers mainly deliver professional patent examiners.

Beside the commercial data providers there are also non-commercial activities such as PatentLens [4], Freepatentsonline [5], or WikiPatents [6]. Patent Lens and freepatentsonline are free patent portals that provide a web-based user interface for searching and reading patents. Both approaches still lack the tight integration of users within a community process for commenting and reviewing patents. WikiPatents on the other hand is a wiki system that contributes to the US patent system by reviewing issued patents and pending patent applications. WikiPatents is related to our approach, but does not use a semantic model – neither for representing patent information nor for representing user annotations. WikiPatents also does not provide advanced visualizations for analyzing large patent sets.

Semantic wikis, such as Semantic MediaWiki [7], will become an important component in a future Semantic Web for adding machine-interpretable content. Semantic MediaWiki provides an extension to the wiki syntax and an enhanced article view to present the interpreted semantic data. It enhances the Wikipedia category concept and introduces typed links, attributes and semantic templates. Semantically annotated wiki pages can be exported in RDF.

In the field of patent visualization the focus has been on document visualization using different dimension reduction methods. The most prominent examples are WebSOM [8] and VxInsight [9]. Visualization techniques also have been applied for patent metadata [10, 11]. Since well designed visualizations can provide aggregated overviews of large patent collections, we believe that it is important to also add visualizations to wiki systems.

3. Semantic Model for Patent Information

For semantically representing different patent information aspects it is useful to interpret patent documents as knowledge objects. Knowledge objects are defined in terms of concepts, properties and relations. Due to performance reasons we use a dual approach. Patent documents are stored in XML ST.36 format [12] in a database and annotations are stored in a separate knowledge base. The knowledge base is realized using the Jena RDF framework [13] in tandem with a relational database backend.

Different patent information aspects, e.g. patent metadata, domain concepts, patent classification, images within patents, the patent structure, etc. are grouped in different ontology modules. Figure 1 gives an overview of the different ontology modules. In the following, we can only give a brief overview of each

module. A more detailed description can be found in [14, 15].

To capture common sense knowledge in patents, the *Suggested Upper Merged Ontology (SUMO)* [16] is used. All other modules are linked to SUMO. One of the reasons why SUMO has been chosen as upper level ontology is that it is linked to the English WordNet [17]. WordNets for the European patent languages English, French and German form the *linguistic ontologies*. Patent genre specific knowledge is encoded by means of a *Patent Upper Level Ontology (PULO)*, which subsumes modules for patent metadata, Structure, Drawings, and Classification. PULO defines patent-specific but domain independent concepts and acts as a bridge between the high-level abstractions of SUMO and the low-level details of the domain ontologies. We distinguish two types of domain ontologies, *core domain ontologies* and *auto domain ontologies*. The first are manually created and comprise a medium number of elaborated concepts, properties, and relations and are linked to WordNet. The latter are automatically created using an ontology learning process based on the core concepts [18]. The proposed ontology framework accounts for a homogeneous semantic representation of patent material merged from different patent services. It also functions as integration platform for user driven annotations in the *user ontologies and folksomies*.

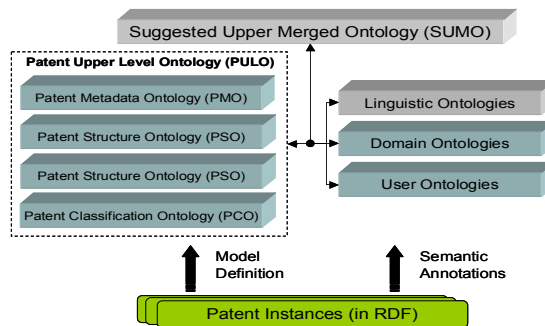


Figure 1: PATExpert Ontology Framework

4. User Requirements

The group of patent information users is rather heterogeneous. However, typical users are patent department employees, patent attorneys, examiners at the patent offices, inventors, or analysts. To get a better understanding to the user requirements, a preliminary study was conducted by the Fraunhofer Patent Center for German Research at the beginning of PATExpert. The presentation of the complete study would be out of the scope of this paper. However table 1 shows the

results of the user study concerning the a) requirements for tools and the b) requirements for team-work.

Question	Yes	No
1. Do you have the necessity to change the layout of the result presentation?	75,0	25,0
2. Do you want to include comments into the results?	68,8	31,2
3. Do you have to exploit the results in cooperation with other persons?	46,2	53,8
4. Do you make the inquiries in cooperation with other persons?	30,8	69,2
5. Do you have the necessity for storing the original results of an inquiry?	82,4	17,6
6. Do you want to store meta data from the relevant patents?	81,3	18,7
7. Would it be helpful to have some kind of tool-tip within the images which can show an explanation of details?	100,0	0,0
8. Some tools are offering the possibility to highlight the hits. Is this helpful for your daily work?	69,3	30,7

Table 1: User answers related to tool usage and team-work

As a direct consequence of the user requirements, the system design has been focused on the integration of visualizations and the possibility of making semantic user annotations.

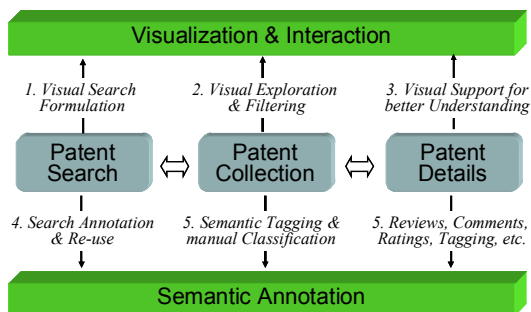


Figure 2: Patent search process support

In general, a patent search consists of three phases (see figure 2). The first phase is the formulation or reformulation of a query. The second phase is the coarse analysis of the result set. The third phase is a detailed analysis of individual patent documents. For each step different visualization and annotation interfaces can help the users in performing its tasks. The aim of the proposed wiki system is to serve as integration platform for different visualization and annotation interfaces as described in the following section.

5. Semantic Patent Wiki

This section describes our approach for combining patent search, patent visualization and patent annotation in a semantic wiki system. Figure 3 gives an overview of the components and the data flow between them. There are three ways to add content to the wiki. First, a user executes a query (step 1) and accepts the results (step 6). Second, the user adds semantic annotations (step 7). Third, the update manager adds newly published patent data to the wiki system in an asynchronous process. In the following we describe a typical search scenario.

The first step is the formulation of a query. A patent index is looked up in order to retrieve the patent numbers of the resulting patent documents (step 2). In PATExpert this step involves four different search engines: a full-text search engine, a semantic search engine for searching in the knowledge base, a metadata search engine and an image search engine. The result of step 2 is a ranked list of patent document numbers.

In step 3 the patent contents are looked up in the cache. In PATExpert the cache is realized as a relational database, which uses the data from the Open Patent Services (OPS) [19] and the European Publication Server [20]. In step 3 also the knowledge base is queried for relevant semantic annotations. The result of step 3 is a patent model. A patent model can be serialized in RDF according to the patent ontologies described in section 3.

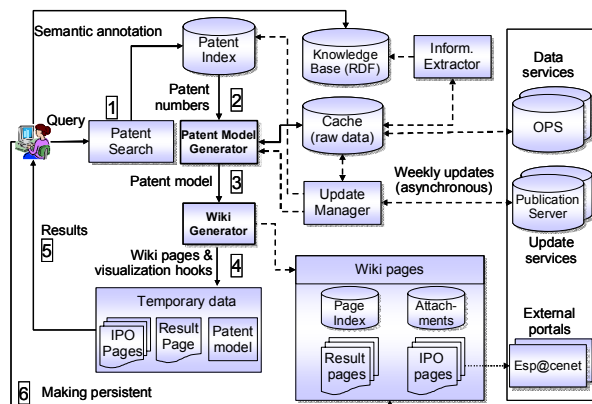


Figure 3: System overview

In step 4 the wiki pages are generated based on the patent model. This generation is customized by scripts. A summarization page is created that contains the resulting patents, the patent model as attachment and visualization hooks. Visualization hooks are wiki plugins that take a patent model as input and create a visualization as output. Visualizations are currently

implemented in two ways: as Java applets on client-side or as images dynamically generated on server-side. The following code gives an example for a client-side applet visualization plug-ins for the JSPWiki [21] system.

```

{{patviz
code='patviz.prefusext.applet.PrefuseXTApplet'
archive='Applets/patviz.jar' width='100%'
height='600' script='Scripts/ipcInit.js'
model='patents.gz'}}

```

The parameters code, archive width and height are the applet parameters. The applet in this example is initialized with a script and a patent model. All model references are made relative to the current wiki page.

In the next step (5) the user can view the results and examine the details. The user can either discard the results and reformulate the query, or add the generated contents to the wiki system in order to persist them (step 6). Persistent pages can be semantically annotated, ranked and reviewed by other users. Adding new pages to the system may also trigger an update of existing pages (e.g. new patent pages could be added to inventor, applicant or category pages). Updates of existing pages trigger RSS notifications for users that have registered, for example to get notified when new patents of specific applicants are added.

User annotations are stored in a global knowledge base and registered users can add annotations. The interface for annotating patent pages currently is realized by customizable form plug-ins. Figure 4 shows a simple patent rating plug-in and its definition.

```

[{{FormOpen form='rateForm'}} ...
| Importance | [{{FormInput type='radio' name='imp' value='high'}}] |
[{{FormInput type='radio' name='imp' value='medium'}}] |
[{{FormInput type='radio' name='imp' value='low'}}] ...
[{{FormClose}}]
[{{FormOutput form='rateForm' handler='RatingRDFPlugin'
populate='handler'}}]

```

The snippet defines a form and attaches it with a *RatingRDFPlugin*. Each form plug-in defines a form handler, which is responsible for adding appropriate statements to the knowledgebase. All user annotations are stored as RDF named graphs [22] with the user id as graph name. This is for example necessary to be able to deal with spam.

6. Patent Visualization Framework

This section describes the patent visualization framework (PatViz). The aim was to provide multiple interactive and coordinated visualizations for various patent information aspects. The visualizations developed so far support the analysis of temporal,

geographic, classificatory, and term distributions of large patent sets. PatViz uses the Prefuse framework [23] and extends it for 3D.

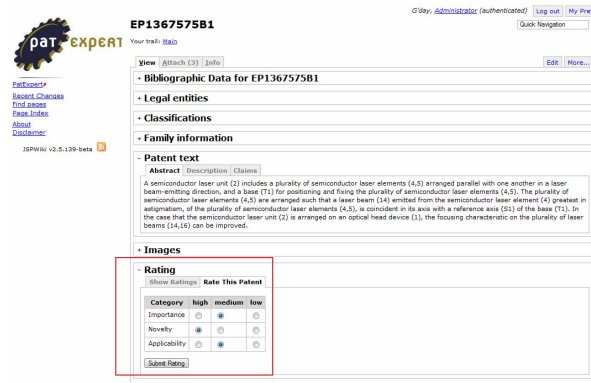


Figure 4: Simple rating plug-in

Input for the visualizations are the patent models attached to wiki pages as described in section 5. Patent models are Java object models that encapsulate the application logic and provide application programming interface. Patent models can be serialized in RDF.

One important design goal was to minimize the loading time for the data transferred to the client as well as the time needed for parsing. We therefore use efficiently compressed binary object streams that are directly deserialized into a Java object model on client-side without any parsing. We are aware that this is specific to the patent domain, since we work with immutable data.

Prefuse originally is a 2D framework. Providing a third dimension offers new possibilities, e.g. for the integration and linking of multiple views in one common 3D space [25, 26]. Therefore new OpenGL rendering components have been developed that re-interpret 2D models in order to create a 3D scene. As a result, original 2D models can be positioned in a 3D space and linked together.

Figure 5 shows the patent visualizations derived from the attached patent model of a result page. The right upper graphic shows the distribution of the patent set according to the International Patent Classification (IPC) using an ordered treemap [24]. The amount of classified patents of a certain IPC category is indicated by color. The user can zoom in and expand sub-categories. The right center graphic shows the patent set as search and filterable table using a provided wiki format. In the right lower visualization the geographical distribution is presented as a colored word map, where the colors indicate the amount of patent in a certain country.

The left upper part shows a combination of different visualizations in a coordinated 3D-view,

where each 2D-visualization is projected on a different wall of a cube. Since all visualization are based on the same model, the selection of items in one part is reflected by filter actions in the other parts (brushing and linking). The left lower visualization shows the extracted key term of the patent set as tag cloud.

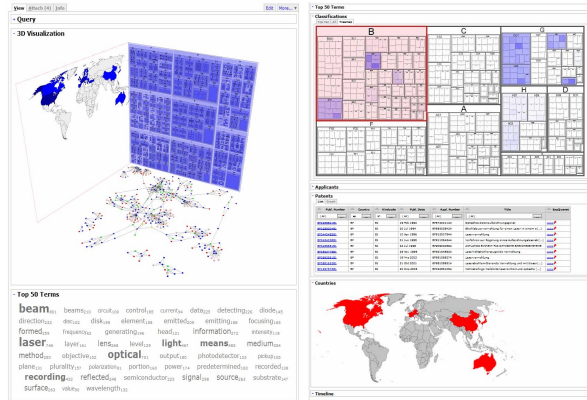


Figure 5: Visualization of search results

7. Discussion

We believe that integrating visualizations into the patent wiki allows for additional analysis of patent sets. Since the focus was to provide client-side interaction and rich visualizations, we used and extended an existing visualization framework and integrated the visualizations into the wiki as Java applets. If applets are not available, there is a second solution for rendering the visualizations as images using the same visualization framework on sever-side. Instead of applets, JavaScript is used on client-side for providing basic interactions, such as hovering or selecting items. For the client-side scripts we use the Mootools [27] framework.

The visualization framework itself is managed as a special wiki page, which has the framework code and the configuration scripts attached. As a simple extension the script pages could be associated to different user roles in order to provide some further level of personalization. For fully personalization the scripts could be associated with the user home page. The latter is not foreseen at the moment.

Another important aspect is privacy. We are aware to the fact that searching for patent information is generally considered as sensitive, because it could give competitors hints about the directions in which a company could go concerning the current and future research. This problem is not dealt with in this paper, but we want to give two possible solutions to this issue.

First, wiki user management functionality can be used to restrict the access to certain pages considered as sensitive. This of cause means to trust the provider, who runs the wiki service. A second solution would be to encrypt the sensitive parts of the information, so that only the owner or a restricted group can access this information. The aspect of partial encryption of RDF graphs has been discussed in [28].

Currently semantic annotations can be made via a wiki syntax extension or via special plug-ins, such as the previously described ranking plug-ins. The annotation interface has simple browsing functionality for PULO and the domain ontologies and allows the selection of concepts defined in these ontologies to formulate semantic annotations. An extension of the ontologies is not possible in the current implementation but is planed for the future.

Concerning the patent data the current limitation is the need of a patent index. Currently the patent index contains about 100.000 patents that have been crawled using OPS as test patent corpus developed in the PATExpert project. The patent domains investigated in PATExpert are restricted to ‘optical recording’ and ‘machine tools’. This limitation will change in the near future because OPS will offer advanced search functionality in its next version. Full-text and metadata search is then possible on the complete patent database of the European Patent Office. Additional semantic search is possible on the patents added to the wiki. By using a semantic representation formalism for patent information, it is expected that other patent services can be better integrated and merged in the future.

8. Conclusions

We presented a new approach for searching, presenting and annotating patent information using semantic wikis. The JSPWiki system has been extended in order to integrate interactive visualizations for patent information. The visualizations currently focus on providing deeper insights into the classificatory, geographical, and temporal distribution of patent sets. These aspects will be extended in the near future to semantic similarities between patents by making use of the automatically and manually extracted semantic annotations.

Our approach differs from typical wiki usage scenarios in the sense that it combines automatic content generation based on patent search activities of the users with user driven semantic annotation of patent information. Currently the aspects of patent rating, linking with prior art, reviews, and general comments and discussions are considered. The content generation is based on a semantic patent model, which is described in terms of a patent ontology.

The aim of the proposed system was to use and extend a modern wiki system for semantic annotation and to integrate patent search functions and to combine it with rich visualizations in order to support the knowledge-intensive tasks of patent search and understanding.

Acknowledgements

The work presented in this paper has been developed within the PATExpert project. PATExpert is funded by the European Commission within its Sixth Framework Programme (FP6 028116).

References

- [1] PATExpert: <http://www.patexpert.org>
- [2] Microformats: <http://microformats.org>
- [3] G. Klyne and J.J. Carroll (ed.): Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 2004
- [4] PatentLens: <http://www.patentlens.net>
- [5] Freepatentsonline: <http://www.freepatentsonline.com/>
- [6] WikiPatents: <http://www.wikipatents.com/>
- [7] M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, and R. Studer: Semantic Wikipedia, Proceedings of the 15th international conference on World Wide Web, ACM Press, 2006
- [8] K. Lagus, S. Kaski, and T. Kohonen: Mining massive document collections by the WEBSOM method, Information Sciences, 163(1), pages 135-156, 2004
- [9] K. Boyack, B. Wylie, and G. Davidson: Domain visualization using VxInsight for science and technology management, Journal of the American Society for Information Science and Technology, 53:9, pages 764-774, 2002,
- [10] C. Chen, and R. Paul: Visualizing a Knowledge Domain's Intellectual Structure, Computer, 34:3, pages 65-71, 2001
- [11] M. Giereth, S. Koch, M. Rotard und T. Ertl: Web Based Visual Exploration of Patent Information, 11th International Conference on Information Visualization, S. 150-155, IEEE Computer Society, 2007
- [12] Word Intellectual Property Organization standards: <http://www.wipo.int/scit/en/standards/standards.htm>
- [13] Jena RDF Framework: <http://jena.sourceforge.net/>
- [14] L. Wanner, S. Brüggemann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhmann, G. Rao, M. Rotard, P. Schoester, L. Serafini und V. Zervaki, "Towards Content-Oriented Patent Document Processing", in: World Patent Information Journal, 30(1):21-33, Elsevier, 2008
- [15] M. Giereth, S. Koch, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, L. Serafini, L. Wanner und T. Ertl, "A Modular Framework for Ontology-based Representation of Patent Information", in: Legal Knowledge and Information Systems - JURIX 2007, Frontiers in Artificial Intelligence and Applications, Vol. 165, S. 49-58, IOS Press, 2007
- [16] I. Niles, and A. Pease: Towards a Standard Upper Ontology, 2nd International Conference on Formal Ontology in Information Systems, 2001
- [17] C. Fellbaum: WordNet: An Electronic Lexical Database MIT Press, 1998
- [18] A. Potrich und E. Pianta: Learning Domain Specific Isa-Relations from the Web, in: 6th International Conference on Language Resources and Evaluation (LREC-2008), 2008
- [19] Open Patent Services: <http://ops.espacenet.com/>
- [20] European Publication Server: <https://publications.european-patent-office.org/>
- [21] JSPWiki: <http://www.jspwiki.org/>
- [22] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler: Named Graphs, Provenance and Trust, Report HPL-2004-57R1, <http://www.hpl.hp.com/techreports/2004/HPL-2004-57.html>, 2004
- [23] J. Heer, S. K. Card, and J. A. Landay. "Prefuse: A Toolkit for Interactive Information Visualization", In: proceedings of CHI 2005, 2005
- [24] B. Shneiderman, M. Wattenberg: Ordered Treemap Layouts, IEEE Symposium on Information Visualization (InfoVis), pages 73-78, 2001
- [25] C. Collins, S. Carpendale: VisLink: Revealing Relationships Amongst Visualizations, IEEE Transactions on Visualization and Computer Graphics, 13(6):1192-1199, 2007
- [26] J. Roberts. State of the Art: Multiple & Coordinated Views in Exploratory Visualization, in 5th Int. Conf. in Coordinated and Multiple Views in Exploratory Visualization, 2007
- [27] Mootools: <http://mootools.net/>
- [28] M. Giereth: On Partial Encryption of RDF-Graphs, In: The Semantic Web - ISWC 2005, Springer, pages 308-322, 2005