Semantic Web 1 (2012) 1–5
IOS Press

Publishing and Interlinking the USPTO Patent Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Amrapali Zaveri [a,*], Mofeed M. Hassan [a], Tariq Yousef [a], Simon Chill [a], Sören Auer [a] and Jens Lehmann [a]

[a] *Universität Leipzig, Institut für Informatik, AKSW, Postfach 100920, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de

Abstract. Patents are widely used to protect intellectual property and a measure of innovation output. Each year, the USPTO grants over $150,000$ patents to individuals and companies all over the world. For instance, there were more than $300,000$ patent grants issued in the US in 2013. However, accessing, searching and analyzing those patents is often still cumbersome and inefficient. To overcome those problems, Google indexes patents and converts them to XML files using OCR techniques. In this article, we take this idea one step further and provide semantically rich, machine-understandable patents in RDF format. This data can be integrated with other data sources in order to further simplify use cases such as trend analysis, structured patent search and exploration and societal progress measurements. We describe the conversion, publishing, interlinking process along with several use cases for the USPTO Linked Patent data.

Keywords: RDF, XML, Patents, USPTO

1. Introduction

A patent is a set of exclusive rights granted to an inventor by a sovereign state for a solution, be it a product or a process, to a particular technological problem [6]. The *United States Patent and Trademark Office* (USPTO)[1] is part of the *US department of Commerce* and grants patents to businesses and inventors for their inventions in addition to registration of products and intellectual property identification. Each year, the USPTO grants over $150,000$ patents to individuals and companies all over the world. As of December 2011, more than 8.7 million patents have been issued and 16 million applications have been received[2].

Patents are a form of intellectual property and a measure of innovation output. They cover a broad range of technologies and are a rich source of information. Patents have the following applications:

- facilitate transmission of knowledge from academia to the industry and its application for industrial purposes [8],
- technological indicator as a sign of transition between science and product, which in turn can be connected to policies of interest such as innovation, economic growth, welfare [7],
- measure the output of research and development, its productivity, structure and development in a particular technology or industry or specific domains,

*Corresponding author. E-mail: zaveri@informatik.uni-leipzig.de.

[1] http://www.uspto.gov/

[2] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm

- track the level of dissemination of knowledge across technology areas, sectors, firms, countries etc.[3],
- statistical indicators of the inventive performance and output of countries, regions, technologies, firms etc.

However, there may be cases when some inventions cannot be patented or a particular country does not incline towards patenting their innovations, which makes it difficult to compare patenting activities across countries. On the other hand, patents, which particularly focus on cross-boarder tie-ups assist in international comparability, indicate the international flow of knowledge as well as funds for research from the inventor country to the applicant countries[4].

The USPTO patents are accepted in electronic form and are filed as PDF documents. Google is indexing these patents after using optical character recognition and making them searchable[5]. A total of 7 million patents dated from 1790 onwards are available. However, the indexing is not perfect and it is cumbersome to search through the PDF documents. Additionally, Google has also made all the patents available for download in XML format, albeit only from the years 2002 to 2012[6]. We converted this bulk of data (spanning 10 years) from XML to RDF conforming to the Linked Data principles [3].

In this article, we first describe the ontology designed for representing patent data in Section 2. The process of the conversion and publishing of the USPTO patent data from XML to RDF is documented in Section 3. Details of the interlinking of the data with other datasets are presented in Section 4. Section 5 portrays a few potential application scenarios and use cases for the converted patent data. A number of related initiatives are discussed in Section 6. Finally, we conclude with lessons learned and future work in Section 7.

2. An Ontology for Representing Patent Data

Every `Patent` is associated with a unique document ID, a title, the kind of patent, country where it was issued and date of publication. Additionally, each patent has an `Applicant`, who applied for the patent. Each applicant's first name, last name, nationality and residence information is provided accompanied with the city, state and country of origin. Moreover, each patent file contains data about other patent `Patents` that have cited this particular patent. These referenced documents have an associated name and category. Figure 1 shows an overview of the ontology underlying the USPTO patent data.

3. Dataset Conversion and Publishing

The USPTO patents full-text data is available for download in XML format from the years 2002 onwards[7]. From the years 1976 to 2001, the data is available in a plain-text format. Each week USPTO releases a zipped file of all patents accepted in that week. Each year ca. $52 - 55$ files are published each one containing about $5,000$ patents.

```
1  PREFIX uspatent:
2  <http://us.patents.aksw.org/patent/>
3  PREFIX uspatent-s:
4  <http://us.patents.aksw.org/schema/>
5
6  uspatent:D0651376
7    rdf:type                    uspatent:Patent.
8    uspatent-s:docNo            "D0651376"@en;
9    uspatent-s:kind             "S1"@en;
10   dcterms:date                "2012-01-03"^^xsd:date;
11   uspatent-s:inventionTitle "Candy holder"@en;
12   uspatent-s:citedBy          uspatent:D11495;
13   uspatent-s:country          uspatent-s:US;
14   uspatent-s:applicant        applicant:D0651376-
         Butts-Cornish .
15
16 uspatent-s:US rdfs:label  "US"@en .
17
18 applicant:D0651376-Butts-Cornish
19    rdf:type                  foaf:Agent;
20    foaf:lastname             "Butts-Cornish"@en;
21    foaf:firstname            "Barbara A."@en;
22    uspatent-s:nationality  "omitted"^^xsd:string;
23    uspatent-s:residence      uspatent-s:US;
24    uspatent-s:city           "Norco"^^xsd:string;
25    uspatent-s:state          "CA"^^xsd:string .
```

Listing 1: Excerpt of the RDF representation of a single patent from the year 2012.

We utilized the XSLT (Extensible Stylesheet Language Transformations) language to transform the XML data to RDF. The XSLT stylesheet contains a collection of rules, instructions and other directives

[3] http://www.oecd.org/sti/msti.htm
[4] http://www.oecd.org/science/inno/21682515.pdf
[5] http://www.google.com/patents
[6] Patents available before 2002 are available in .txt format.
[7] http://www.google.com/googlebooks/uspto-patents-grants-text.html

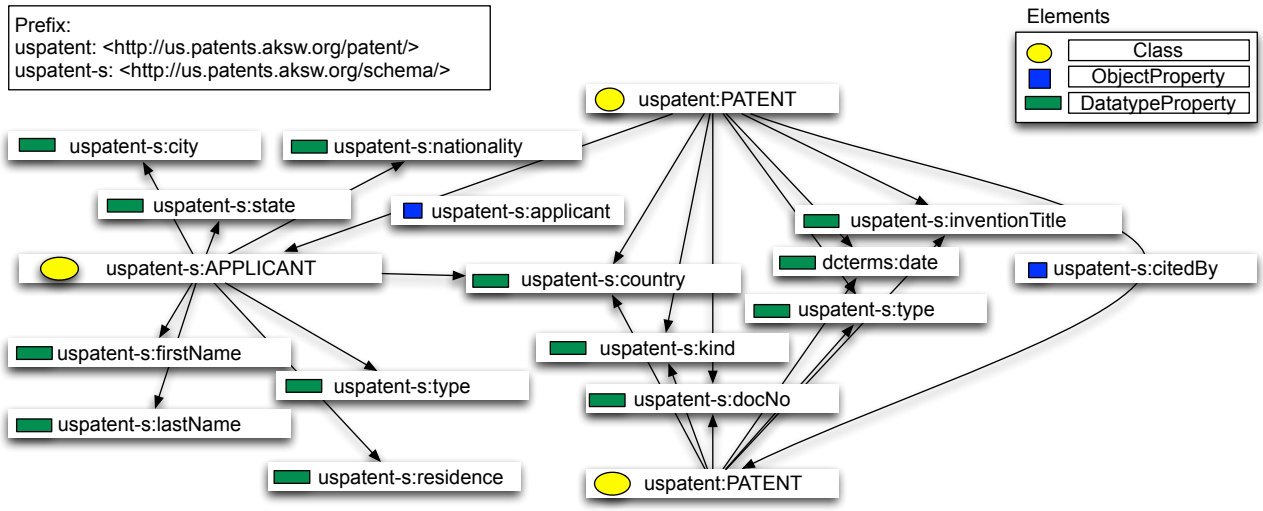

Fig. 1. Depiction of the ontology underlying the patent data.

that guide the processor in the production of the output document.

In order to assist with the conversion, we developed a semi-automated tool[8] enabling a user to choose (i) an input XML patent file, (ii) the corresponding XSLT file and (iii) an output RDF file name and location. An example of an input patent in XML, the XSL transformation and the corresponding RDF output is shown in Figure 2. For every input XML file, the user obtains a corresponding output RDF file using turtle notation. In order to convert one year including all the $52 - 55$ files, the tool requires about one hour. After we converted the patent data from the years 2002 - 2014, a total of 150 GB RDF data was produced amounting to c.a. 187 millions of triples. Due to the changes in the structure of the XML files for the years 2002 to 2004, 2005 to 2012 and 2013 onwards, it was necessary to create three different XSL transformations. It was observed that the years from 2002 to 2004 had missing information such as country, category and name for citations as well as applicant information. An example of a single patent is displayed in Listing 1 and an example patent RDF file from the year 2013 can be downloaded here[9]. In order to keep the data up-to-date, we created an automated script[10], which detects whether new patent XML files are uploaded on the google patents storage site[11] and converts them to the corresponding turtle format.

After converting the data, we published it as Linked Data using the *OntoWiki* platform [2]. OntoWiki not only allows the publishing and maintenance of Linked Data but also provides a SPARQL endpoint for the dataset in combination with Virtuoso[12] as the storage back-end for the RDF data. Additionally, it is also possible to browse the data with the HTML output of OntoWiki. Moreover, publishing with OntoWiki makes the URIs dereferenceable. Thus, the published USPTO Linked Patent data is available at `http://us.patents.aksw.org/`[13]. Links to the dataset, SPARQL endpoint, VoID file and the DataHub entry along with information on the version date, number and licensing are listed in Table 1.

4. Dataset Interlinking

Abiding by the fourth Linked Data principle of including links to other related things (using their URIs) when publishing data on the Web, we identified several target links from external datasets. In particular, we interlinked the countries between USPTO and DB-

[8] Available at `https://github.com/chillSen/USPTO-XML2RDF-Tool`

[9] `http://us.patents.aksw.org/export/2013.zip` (c.a.359M)

[10] `https://github.com/chillSen/USPTO-XML2RDF-Tool`

[11] `http://www.google.com/googlebooks/uspto-patents-grants-text.html`

[12] `http://virtuoso.openlinksw.com/`

[13] For example, see resource `http://us.patents.aksw.org/patent/08189687`

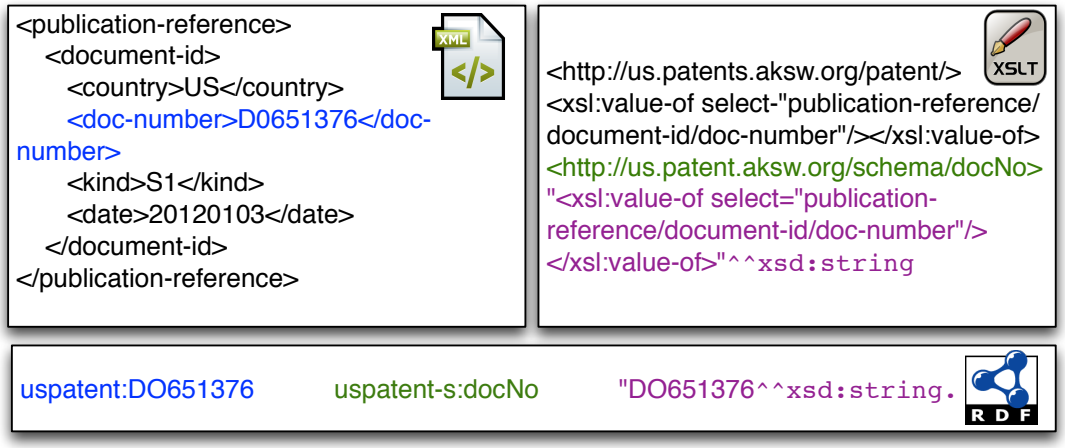

Fig. 2. Transformation of the input XML (left) using the XSLT configuration (right) to the output RDF (bottom).

Table 1
Technical details of the USPTO RDF dataset.

Namespace	`http://us.patents.aksw.org/`
Total no. of triples	c.a. 187 million
SPARQL endpoint	`http://us.patents.aksw.org/sparql`
Graph name, Interlinks graph name	`http://uspatents.aksw.org/`, `http://uspatents.interlinks.aksw.org/`
Version date and number	June-2104, 1.0
Licensing	cc-by
VoID File	`http://us.patents.aksw.org/export/void.ttl`
DataHub entry	`http://datahub.io/en/dataset/linked-uspto-patent-data`
Homepage	`http://www.aksw.org/Projects/USPatents.html`

pedia, World Bank as well as EU Patents. The links between the USPTO and the DBpedia as well as WorldBank countries allows one to explore and compare related information about that country such as amount of investment by the government in patent related activities versus the GDP of that country, for instance. The links between USPTO and EU Patents ensures a greater coverage of the patents, which helps users (in this case potential investors or companies) get an idea of the market trends or competitors in European countries.

We performed the interlinking using the LIMES [5] framework, which allows a user to specify the source and target properties; metric; relation type as well as the threshold based on which links are discovered between datasets. Since the amount of USPTO data is large, the time taken for LIMES to discover the links would be very large. Also, due to the incapability of LIMES to match the country codes to the country URIs in one query, we first extracted the country URI and abbreviations from the datasets and used this file[14](LIMES also allows CSV files as input instead of SPARQL endpoints) as input to carry out the interlinking. The threshold was set to 0.8 and there were no links obtained below this threshold. Thus, the links obtained are considered to be exact matches and thus of high quality. Random manual checks were additionally done for a subset of the links to ensure that the links obtained were indeed correct. The details of the number of source and target instances as well as the number of accepted links along with the precision and recall for each is recorded in Table 2.

5. Application Scenarios and Potential Third-party Use-Cases

In this section, we provide selected application scenarios and potential third-party use cases for the Linked Patent data.

[14] A CSV file with two columns: country URI and abbreviation.

Table 2

Number of interlinks obtained between USPTO and DBpedia, WorldBank and EU Patents for countries.

Links between	Link type	Source dataset	Source Instances	Target Dataset	Target Instances	No. of Links	Precision	Recall
Countries	owl:sameAs	USPTO	215	DBpedia	245	215	1	0.87
Countries	owl:sameAs	USPTO	215	World Bank	214	192	1	0.89
Countries	owl:sameAs	USPTO	215	EUPatents	23	23	1	1

Measuring societal progress. The patent dataset, when combined with indicators from the World Bank dataset, for instance, can be used to measure specific societal progress indicators.

```
SELECT ?countryP ?noofpatents ?year ?obsValue
WHERE
{ GRAPH g-indicators: { ?observationURI
 property:indicator   indicator:GB.XPD.RSDV.GD.ZS;
 sdmx-dimension:refArea  ?refAreaURI;
 sdmx-measure:obsValue   ?obsValue. }
GRAPH g-meta:
{  ?refAreaURI  a               dbo:Country .
   ?refAreaURI  skos:prefLabel  ?countryWB. }
SERVICE <http://us.patents.aksw.org/sparql>
{ SELECT COUNT(DISTINCT(?s)) AS ?noofpatents ?
    countryPT ?year
WHERE {
?s          patent:country      ?country.
?country    rdfs:label          ?countryPT.
?countryPT  owl:sameAs          ?countryWB.
?s          dcterms:date        ?year. } } }
```

Listing 2: SPARQL query retrieving the number of patents per country and the R&D expenditure as a % of GDP (from the WorldBank dataset).

For example, the total number of patents per country can be used as an indicator for innovation, which in turn can serve as a proxy measure for the output of R&D in the form of inventions in that country. This number can then be compared with the R&D expenditure (% of GDP) for each country, which can be derived from the World Bank dataset.

Listing 2 shows the SPARQL query, which retrieves the country name, total patent count for that country (from the patent dataset) with the amount of R&D expenditure for that country (from the WorldBank dataset). This comparison can thus be used to measure whether the amount invested by a government in R&D is proportional to the innovation churning out of that investment. This use case is made possible due to the uniform structure of the datasets (i.e. RDF) allowing seamless integration, which would require substantial work with XML data.

Trend analysis. Another advantage of having all the patents in the RDF format is that one can analyze the patents over time by querying for the number of patents in each year. This analysis is possible since the data is stored in a uniform format and thus can be easily aggregated. Using this information, one can measure the growth in number and type of patents over these years, thus allowing governments to analyze the growth in innovation and measure the effect of the investment in research.

```
SELECT COUNT(DISTINCT (?patent)) ?year
FROM <http://uspatents.aksw.org/>
WHERE { ?patent dcterms:date ?year. }
GROUP BY ?year
```

Listing 3: SPARQL query retrieving the total number of patents per year thus allowing to perform trend analysis.

Most often, companies also need reports about trends in a particular field over the past 10 years. Moreover, this patent information and analysis can be helpful in identifying new directions of technology that is perhaps currently lacking by querying for patents belonging to a particular field (e.g. biomedical).

This type of analysis would be cumbersome and error-prone when dealing with XML data, but the RDF data makes this kind of analysis faster and easier for the user. Listing 3 shows the SPARQL query for retrieving the total number of patents per year. This can of course be extended to include particular countries or particular areas of interest. Figure 3 shows the trend in total number of patents over the years 2000 to 2013.

Data exploration. Patent analysts are specifically employed to look into a particular technology development for a company interested in investing in that field. The analysts have to *manually* search through the plethora of information to find patents, which match their area of interest. Thus, they are faced with a big

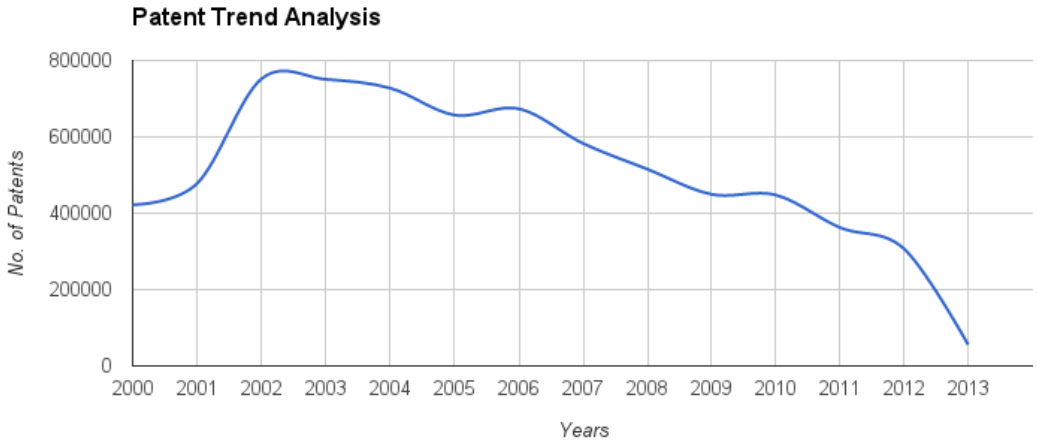

Fig. 3. Trend analysis: showing the growth in total number of patents over the years 2000 to 2013.

challenge to extract valuable information from a huge dataset. Moreover, executing tasks such as carrying out successive comparison and refinements to narrow down the search results to the set of related patents is a time consuming and error-prone task. RExplorator [9] is a tool, which allows exploration of (semi) structured data repositories.

Currently, RExplorator is being tested to explore the RDF patent data. With the help of this tool, tasks such as searching, refining, comparing and grouping results will be streamlined to help the analysts perform their analysis effectively. With data available as RDF, carrying out search over the data will be efficient as one can query for specific types of patents, for specific countries or years by means of a single SPARQL query (as opposed to searching through the XML data).

6. Related Initiatives

There are three different areas of related work that can be identified: (i) dataset conversion, (ii) dataset publishing and (iii) data search, which we briefly discuss in this section.

Dataset conversion. There are a number of tools, which contribute towards generalizing the conversion of XML to RDF:

- AstroGrid-D's XML to RDF via XSLT transformation [1]
- Using XQuery to transform XML to RDF[15]
- Krextor RDF Extractor, which is an extensible XSLT-based framework for extracting RDF from XML[16]
- XSPARQL, which merges both XQuery and SPARQL for mapping XML and RDF
- LODRefine, which is the LOD enabled version of Google Refine and can take XML files as input (among other formats) and transform it to RDF with the help of the RDF extension

Dataset publishing. Some of the freely available patent datasets have already been converted to Linked Data, which are discussed in this section.

European patents. The European Patent Office (EPO) provides access to millions of European patent documents. Data for the last six weeks can be obtained for free through the data portal of the EPO whereas older documents need to be obtained via a DVD on a paid basis. This dataset was converted to RDF and is available at `http://epo.publicdata.eu/`. The dataset contains more than 10 million triples and links to DBpedia and Eurostat.

UK patents. The patents belonging to the UK government have also been converted and published as Linked Data. However, it has last been updated only

[15] `http://www.ibm.com/developerworks/xml/tutorials/x-xqrdf/section2.html`

[16] `http://trac.kwarc.info/krextor/`

in 2010 and is available only as a dump[17] or via a SPARQL endpoint[18].

Data search. Additionally, patent data is already available for search via a number of websites or web portals, as discussed in this section.

Google Patent Search. All patents available through Google Patent Search come from USPTO. Patents issued in the US are public domain government information, and images of the entire database of U.S. patents are readily available online via the USPTO website. Google Patent Search covers the entire collection – 7 million – of issued patents and millions of patent applications made available by the USPTO from patents issued in the 1790s through those most recently issued in the past few months. Additionally, Google introduced a *Prior Art Finder* in order to determine the novelty of a patent by looking for similar patents that existed at the time the patent was filed[19]. However, the data is available as only text or XML formats and is not interlinked with other datasets to discover interesting information.

WIPO IP Statistics Data Center. The World Intellectual Property Organization (WIPO) Intellectual Property (IP) Statistics Data Center enables access to the WIPO's statistical data about patents[20]. For example, basic statistical information such as the total number of patent applications, patent grants, international applications, patent publications by technology and complex statistics such as resident applications per 100 billion USD GDP by origin etc. These categories can be supported by additional filtering based on the report type, which helps users specify the territory of protection ("By office") or the owner of the protection ("By origin"). Additionally, filtering by the year and country is also possible.

The Lens. The Lens[21] is a platform to explore and understand the patent system. The platform allows the user to make informed decisions, explore areas of innovation in which they can be creative and make wise investments based on complete information about patents. Information is pulled in from 90 patent jurisdictions worldwide. However, as reported, the data is not always in a machine-readable format making it difficult to search and analyze them [4].

[17] http://source.data.gov.uk/data/patents/bis-research-explorer/2010-03-04/patents.data.gov.uk.nt

[18] http://openuplabs.tso.co.uk/sparql/gov-patents#

[19] http://googleresearch.blogspot.de/2012/08/improving-google-patents-with-european.html

[20] http://ipstatsdb.wipo.org/ipstatv2/ipstats/patentsSearch

[21] http://www.lens.org/

7. Conclusion and Future Work

In conclusion, by providing the USPTO patent data as Linked Data and interlinking it with other datasets, it is possible to not only to search efficiently over the vast amount of information but also use the data in interesting application scenarios in combination with other datasets. However, we did encounter a number of limitations during the conversion such as (i) large amount of data, which affects the efficiency of SPARQL queries and search; (ii) different DTDs from the years 2005 to 2014, thus requiring different XSLT files generations for the conversion; (iii) not all data is modeled such as names of agents and examiners since we chose only the most relevant information from the patents to avoid overload during the conversion. Also, the figures, which are available in the PDF file are not available in the RDF format.

As future work, we plan to continue converting the newly data available and keep the dataset up-to-date using our automated script. Moreover, we intend to look into analyzing the structure and converting the data from the years prior to 2002.

References

[1] F. Breitling. A standard transformation from XML to RDF via XSLT. *Astronomical Notes*, 330(7):755–760, 2009.

[2] P. Frischmuth, M. Martin, S. Tramp, T. Riechert, and S. Auer. OntoWiki - An Authoring, Publication and Visualization Interface for the Data Web. *Semantic Web Journal*, 2014.

[3] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Morgan and Claypool, 2011.

[4] T. Lens. The patent bargain. *Nature*, 504:187–188, 2013.

[5] A.-C. Ngonga Ngomo and S. Auer. LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *Proceedings of IJCAI*, 2011.

[6] W. Publication. *WIPO Intellectual Property Handbook: Policy, Law and Use. Chapter 2: Fields of Intellectual Property Protection*. World Intellectual Property Organization, 2004.

[7] G. d. Rassenfosse and B. van Pottelsberghe de la Potterie. A Policy Insight into the R&D-Patent Relationship. *Research Policy*, pages 779–792, 2009.

[8] T. Reiss and D. I. Lacasa. Benchmarking national biotechnology policy across Europe: a systems approach using quantitative and qualitative indicators. *Research Evaluation*, 16(4):331 – 339, December 2007.

[9] A. S., S. D., and B. S. Experimenting with Explorator: a Direct Manipulation Generic RDF Browser and Querying Tool. *Visual Interfaces to the Social and the Semantic Web*, 2009.