European Patent Information and the CPC Taxonomy as Linked Open Data

Martin Kracker [0000-0002-9720-7264]

European Patent Office, Rennweg 12, 1030 Vienna, Austria
mkracker@epo.org

Abstract. Patent data is a valuable source of business, technical and legal information. The European patent office has recently begun the regular publication of bibliographic information of European patents, including the full CPC taxonomy, in linked data format with an open standard license. Data scientists, commercial providers, web developers etc. can now browse this data set, run SPARQL queries or bulk download the full data set. Linked data significantly reduces the effort of combining heterogeneous data from multiple sources and domains.

Keywords: Patent Information, Technical Classification, Linked Data, Open Data.

1 Introduction

Outside the patent community, patents are an often underestimated data source. Patent information (PI) contains technical, business and legal information. It can serve as a proxy for innovation and it may be used for technical or market intelligence, competitor watch, econometric studies and many other purposes.

With nearly 7 000 staff, the European Patent Office (EPO) is one of the largest public service institutions in Europe. Through the EPO's centralized patent granting procedure, inventors are able to obtain high-quality patent protection in up to 43 countries, covering a market of some 700 million people.

The EPO publishes the European Patent (EP) applications it receives and the patents it grants, currently amounting to nearly 200 000 publications per year.

The EPO disseminates PI to the public according to its four quality criteria: timeliness, completeness, accuracy and usability for many purposes. To fulfil the usability criterion, the EPO distributes its PI in a variety of tools and formats: Online databases, web services and as bulk data[1]. In April 2018 a new flavor of comprehensive EP patent data called *Linked open EP data* has been made available.

[1] www.epo.org/patent-information

2 Related work

Earlier academic work was conducted to define linked data (LD) ontologies for patents [1] and to demonstrate the interlinking between patents and other linked data sets [2]. Few of these data sets are still maintained [3]. The first – and until recently the only - publication of patent information as linked open data (LOD) by a patent office was issued by Korea in 2015[2].

The presented data set from the EPO distinguishes itself by being based on a detailed patent ontology modelled with the expertise of a large patent office. The principal design goal was to maintain a broad scope so it could be used as a blueprint by other patent offices. Like the official patent publications, the represented LOD set will be updated weekly by the EPO who is the competent patent authority.

3 Linked Open EP Data

Linked data is a set of best practices on how to publish structured data. It builds on a set of web technologies which are defined by the W3C[3], like RDF and SPARQL. Data published as linked data facilitates the merging of data sets, including independent information from other sources or organizations.

Linked open EP data is a service of the EPO comprising two published EPO data sets:

- Bibliographic data (i.e. the meta-data) of all EP applications and their publications including the first publications from 1978. This data set is updated weekly.
- The Cooperative Patent Classification (CPC)[4]. The data set is updated with every new CPC version.

In total, the data set contains more than 600 million triples. The target user group for this product goes beyond the typical patent expert and includes web developers and data scientists.

3.1 Content

EP Bibliographic Data.

This data set contains the bibliographic data of EP applications and publications and their related applications and publications of other patent authorities. Bibliographic data on EP patents is quite comprehensive, whereas bibliographic data on non-EP patents is limited to basic patent identification information.

The main resources and their relationships are:

- About 3.2 million EP applications and their earlier filed applications, like priorities or international applications

[2] lod.kipo.kr

[3] www.w3.org/standards/semanticweb/data

[4] www.cooperativepatentclassification.org

- About 19 million non-EP applications which are in the same family as an EP application. Patents which are in the same family cover the same invention, but seek protection in different jurisdictions.
- About 5.5 million EP publications with title, abstract, applicants, inventors and legal representatives
- Technical classifications: CPC and IPC (International Patent Classification[5])
- Citations to patents and non-patent literature, like scientific papers, journals, database entries or web pages
- Links to the full text representation of EP publications, in PDF, XML and HTML format within the EPO's European Patent Server[6]

Cooperative Patent Classification (CPC).

The CPC is a taxonomy with about 250 000 terms which cover the complete technical domain. Its terms are structured according to a strict hierarchy by a narrower / broader relationship. Like the older, less detailed but globally used IPC classification scheme, it is used to classify patents to facilitate patent searches.

3.2 Ontologies Used

Where applicable, linked open EP data makes use of established ontologies, like `rdf/rdfs`, `skos`, `vcard` and `dcterms`. Where necessary, new ontologies have been defined with the goal to allow easy re-usability by other patent authorities.

- The `patent` ontology currently contains 11 classes and 43 properties
- `cpc` and `ipc` model the technical classifications
- `st3` corresponds to WIPO standard ST.3 [7], which represents states and intergovernmental organizations.

4 Access, Bulk Download, API and SPARQL Endpoint

The *linked open EP data* service can be directly accessed at `http://data.epo.org/linked-data` . General information, a user guide and useful resource documentation can be found on the EPO's home page at `http://epo.org/linked-data`.

A download page[8] allows the bulk download of all published data in N-Triples format. The ontologies files are provided in Turtle format.
The API not only resolves URIs, it also supports the Linked Data API (LDA) specification[9] which permits adding parameters to the URI to

[5] www.wipo.int/classifications/ipc/en/

[6] www.epo.org/publication-server

[7] www.wipo.int/export/sites/www/standards/en/pdf/03-03-01.pdf

[8] https://data.epo.org/linked-data/download/

- filter resources
- sort the output and define pagination and to limit the output by pages
- define the output format and apply predefined views

A simple browser displays the retrieved data in a tabular style for easy data and metadata exploration by human users. Clicking on the displayed resources retrieves the related resources. The format of the display of the retrieved data can also be adapted.

Using curl or similar tools, any machine can send requests to the SPARQL endpoint[10]. A query window[11] for interactive use of SPARQL is also provided.

5 Open License and Fair Use Policy

The data is covered by the open standard CC-BY 4.0, which allows the use of the data without costs or registration. Attribution is required.

The use of the additional services, like API and the SPARQL endpoint is governed by a fair use policy as these services are provided for occasional use only. They allow consulting reference data, exploring the data and trying out ideas on a small scale. They are not designed to provide a backend database service for production applications. For these uses, the data must be downloaded in bulk.

6 Conclusion

Patent information contains business, technical and legal information which can add value to many domains outside the patent community. Patent offices, which decades ago published data solely on paper, now typically disseminate patent information in csv or XML format. Going one step further to an even more open data format [4] the EPO began publishing their patent data as linked data in RDF.

We hope that other patent offices will follow suit and create a new, much more flexible and powerful infrastructure for patent information.

References

1. Giereth, Mark Oliver: An Architecture for Visual Patent Analysis. PhD Thesis, Faculty of Computer Science, Electrical Engineering and Information Technology, University of Stuttgart, 2012.
2. Zaveri A., et. al.: Publishing and Interlinking the USPTO Patent Data. Semantic Web Journal (2014)
3. USPTO Patent data from AKSW / University of Leipzig, `https://datahub.io/dataset/linked-uspto-patent-data`
4. 5star Open Data, `5stardata.info/en/`

[9] https://github.com/UKGovLD/linked-data-api/blob/wiki/Specification.md

[10] https://data.epo.org/linked-data/query

[11] https://data.epo.org/linked-data/sparql.html