Executive abstract

Linked data (LD) is a more powerful and flexible method of publishing and accessing data bases relying on international recommendations issued by the World Wide Web Consortium (W3C). The adoption of the LD paradigm facilitates the integration of different data bases and the reuse of vocabularies and knowledge already available as Linked Data.

Linked data of course can be applied to patents too. After some experiments and demos, who in last years exposed as EPO, KIPO, USPTO patents as Linked Open Data, since April 2018 the EPO launched its LOD as a true product, since it is updated weekly.

Linked data in patents will simplify the integration of data provided by different patent offices, and with other scientific and business databases as well. The widespread adoption of patent linked data by info providers and solution integrators is a move to encourage, since it will increase the efficiency of patent analysts day-by-day activities and will extend as well their strategic role in the enterprise. Moreover, patent data could become more used in Competitive intelligence applications.

Table of Contents

Linked data: what and why

Linked Data (LD) is a more powerful and flexible technology of publishing and accessing data bases, backed by a solid background of international standards, carried out by the Word Wide Web Consortium (W3C).

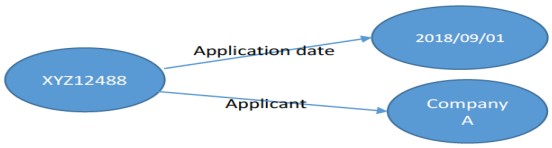
The core idea of Linked Data is to represent a data base as a graph of relationships between entities. Data-bases today are more usually represented as tables, even combined together, as in the relational model. In any case, it is not difficult to transform a table into a graph, which is a more flexible representation. If you have for example this table chunk:

Document id	*Application date*	*Applicant*	*----------*	*---------*
XYZ12488	*2018/09/01*	*Company A*	*----------*	*----------*
---------	*----------*	*----------*	*----------*	*----------*

any cell in the table can also be represented as a triple of values:

a) *Subject (the heading of a row, as XYZ12488): represented as a graph node.*
b) *Predicate (the heading of a column, as application date): represented as a graph edge.*
c) *Object (the cell value, as Company A): represented as a graph node.*

Hence these two cells in the table can also be represented by this graph chunk:



If we apply this transformation to any cell, the table is transformed into a labeled graph, which includes nodes (subjects, objects) and connections (predicates). The graph can be represented on the web, by suitably extending the usual web of documents. The most notable extension of the web of documents is to represent the meanings of links, e.g. "applicant", "application date" and so.

Last, but not least, different kind of Linked Data are typically interlinked together to integrate different kind of databases and to share common vocabularies. Linked Data (LD) are accessed at specific web addresses (end-point). Depending from the end-point, Linked Data can be open or restricted to you and to your business partners. Hence, whilst Linked Data emerged first as Linked Open Data, it is misleading to assume that Linked Data have to be necessarily open as well.

Linked data of course have to be queried: to do that, the W3C has defined the SPARQL query language, the last version of which has been released on 2013. SPARQL allows:

- *to facilitate the integration of different data bases, independently from their structure;*
- *to facilitate the identification of eventual missing or bad data in the original data bases;*
- *to facilitate the expression of more complex queries;*
- *to implement more efficiently advanced applications on the top of your data.*

Just to make an example, imagine that you have identify all patents citing the patent X, directly or indirectly, i.e. through a chain of citations.

The SPARQL query will be (with some simplifications for reasons of clarity):

```
SELECT DISTINCT ?s
WHERE { ?s  cites+  patentX. }
```

- *The "WHERE" clause identifies the subgraph in which the patentX is the object: it is the patent you know;*
- *the predicate is "cites": it includes in this case the recursion symbol +, since we are interested also to indirect citations;*
- *the "?s" are the unknown subjects to be identified;*
- *the "SELECT" clause produces the output of the initially unknown subjects; the clause DISTINCT eliminates eventual duplicated results.*

Linked data available so far in patents

At EPOPIC Kilkenny (2011) an invited speech by Nigel Shadbolt presented the results achieved and/or achievable with linked open data (LOD) in different fields. Since that, Open Data and Linked Open Data entered as a discussion topic in patent conferences.

Some LOD demo were released from 2013 to 2016, typically including a patent set sample, as from KIPO, USPTO and EPO.

In April 2018 the EPO launched its product-level solution LOD, regularly updated (weekly), at https://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1, *which provides also access to the user manual and to the support forum. This is one of the solutions offered by EPO to those using directly EPO data, more typically integrators, available at:* https://www.epo.org/searching-for-patents/data.html. *The solution is provided under the Creative Common Attribution 4.0 License, hence the integrator is free to share and adapt these data, giving proper credit to EPO and without adding restrictions.*

The LOD EPO data base includes:

- *EPO application and publication bibliographic information, including text, weekly updated;*
- *the CPCs tree;*
- *an EPO defined vocabulary of patent concepts (e.g. what is an application) and their relationships;*
- *The data base does not include information about the legal status.*

The EPO LOD entry panel is at https://data.epo.org/linked-data/. *The solution can be accessed in order to download the data base in your system, by accessing* https://data.epo.org/linked-data/download/ *or to make directly a query, by accessing the SPARQL end point, at* https://data.epo.org/linked-data/sparql.html. *In both cases the solution is free. The SPARQL end point has to be used according to a fair use policy; results are obtained according a best effort policy, since the list of results is limited to results obtained in the first minute after the query. Hence the SPARQL end point is to for casual users and experiments. For production-level applications, the LOD data base has to be transferred first to the server of the application integrator which have to assure the suitable performance.*

But LODs of other patent offices are evolving too:

- *KIPO presently provides LOD as a beta: which can be accessed at* http://lod.kipo.kr/, *whilst its SPARQL endpoint has the address* http://lod.kipo.kr/data/sparql *;*
- *USPTO LOD data are available as a demo, at the address* https://old.datahub.io/dataset/linked-uspto-patent-data , *whilst its SPARQL endpoint has the address* http://us.patents.aksw.org/sparql .

Linked data in patents: opportunities and challenges

The adoption of linked data will benefit different layers of a patent informatics solution, as summarized in the following table:

Where (layer)	*Why*
Data base	*To facilitate the verification and cleaning of a data base.*
Data bases integration	*To facilitate the integration of different data bases, as a) patent data bases provided by different patent offices b)*

	other data bases complementary to patents, as scientific publications and business info
Query expression	*To express more powerful queries, as enabled by SPARQL*
Results analysis	*To facilitate the inclusion of more intelligent applications to identify hidden correlations, to filter the most promising results*

Hence the adoption of Linked Data is a win / win opportunity for all players, as patent offices, application integrators and patent professionals.

For patent offices which develop their own data bases, the adoption of Linked Data is suggested even as an internal best-pratice for having a better control of the quality of their data. Moreover, the adoption of this standard form facilitates the distribution of their data to data bases and application integrators and also to other patent offices, when required.

Data bases and application integrators on their side will be facilitated in the integration of patent data bases provided by different patent offices, of compliant to LD standards, and of other patent complementary data bases, with scientific and business data. Moreover, the solution they can provide could support more complex queries and could implement more efficiently advanced applications on the top of your data.

Given these technical evolutions, professional users and their colleagues and managers could rely on more powerful queries and more powerful applications, to increase the efficiency of patent analysts day-by-day activities, and can rely on more data integrated to patent data, as scientific and business data, which can add insight to patent information. On the other side, patent data could become more used in Competitive intelligence applications. Hence the widespread adoption of Linked Data is a trend to encourage.

We are also aware of some possible challenges. Probably the most important challenge against Linked Data is the misleading assumption that Linked Data themselves have to be open as well: this wrong assumption can unduly limit the business model of data providers.