Linked Patent Data: opportunities and challenges for patent professionals

CEPIUG Conference
Milano, 9 - 11 September 2018

Alberto.Ciaramella @intellisemantic.com

This presentation and its objectives

- This presentation is
  - a lean tutorial about linked data
  - a status of the art about linked data in patents
  - a position paper
    - to identify how today technology can cover patent professionals needs
    - to encourage patent offices to extend the adoption of this paradigm in patent information they publish

a network effect can occur

Index

- The framework:
  - what are Linked Data?
  - why and when to use them?
- A Linked Data technical introduction
- Linked Data in patents
  - EPO
  - Others
- Opportunities and challenges in patents
- Conclusions

intellisemantic

A framework and a technical introduction

What are Linked Data?

- Linked Data (LD) is a new more powerful and flexible technology of publishing and accessing data bases
- Linked data is backed by a solid background of **international standards**, developed by the Word Wide Web Consortium (W3C)
- “Linked data” were originally proposed by Berners Lee (2006), as Linked Open Data (LOD), but it is now well recognized that this is only a specific case of use:
  - Linked data **can also be private**, of course.
- More recently, the focus of Linked Data research moved to **engineering aspects** as privacy, security, efficiency, quality of data
- Linked data is one of the enabling technologies for **Big Data** applications, more specifically to solve the challenge of the **variety of data**

Who, where, why linked data (e.g. in patents)

Who	Where (layer)	Why
Data base supplier	Data base	To facilitate the verification and cleaning of a data base
Application developer	Data bases integration	To facilitate the integration of different data bases (different patent offices, non patent literature ..)
Application developer / users	Query expression	To express more powerful queries
Application developer /users	Results analysis	To facilitate the inclusion of more intelligent applications

Evolutions of Linked Data in patents

- At EPOPIC Kilkenny (2011) an invited speech by Nigel Shadbolt presented the results achieved and/or achievable with open data 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
  - **notable demos** by KIPO, USPTO and EPO
- in April 2018 the EPO launched its **product-level solution** LOD, regularly updated (weekly)
  - this seems to be a **turning point**!

A step back - Data base models: from tables to graphs

- A well established and used model in data bases is the relational model, defined in the ‘70s
- in these systems (RDBMS), data are represented as a set of related tables. As an example:
  - a first table associates to any patent its data, including applicant and inventor
  - other related tables include further details about applicants and inventors
- this paradigm (*table-based*) is efficient, but, once designed, it is more difficult to evolve and scale
- in any case, it is straightforward to **convert tables to graphs** (see next slide), and this adds a lot of flexibility: this conversion is the basic concept of **Linked Data**

From tables to (triple) graphs: an example

If we have this table chunk:

Patent id	Application date	applicant	------------	----------
XYZ12488	2018/09/01	Company A	-----------	-----------
---------	------------	------------	----------	------------

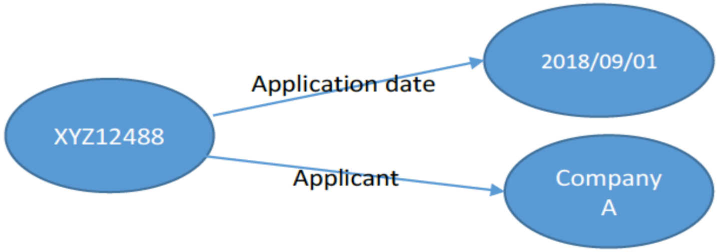
Any **cell** in the table can be represented as a **triple of values**:

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 trasformation to any cell. the table is transformed to a labeled graph, which includes **nodes (subjects, objects)** and **connections (predicates)**

From triple graphs to LD on the web

Linked Data (LD) **implement on the net** *(internet or intranet)* the graph representation just described. To do that:

- **reuse** well known internet standards
  - to name things (with URI)
  - to access them (with HTTP)
- **define** new kind of information, i.e. triples
  - triples are not required in the “web of documents”
- **include external references** to easily reuse knowledge already defined in other graphs, as:
  - “same level” complementary data, e.g. ***data from scientific papers*** to integrate patent data
  - “higher level” common vocabularies, e.g. the ***definition of author***.

SPARQL to query LD: what and why

- what is
    - SPARQL is the **standard** query language for LD
    - it means “SPARQL Protocol and RDF Query Language”
    - SPARQL has been defined by W3C
        - the last version 1.1 has being released on 2013
- why to use
    - 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 implement more efficiently advanced applications on the top of your data
- how to use
    - a LD is typically provided by a SPARQL **end point**, i.e. an internet address which accepts SPARQL queries

SPARQL: the query

- The query is represented by a **subgraph** on the whole graph, having defined **fixed point(s)** in the whole graph
    - e.g. the patent applicant
- Different kind of queries can be represented, from “usual” queries to more involved, still interesting queries, as:
    - example 1: identify all patents of the applicant x published by the authority y in the years zi to zn
        - *this is a more usual example*
    - example 2: identify all patents which are co-cited by at least a member of the patent family A and by at least a member of the patent family B
        - *this is a little more advanced example*

SPARQL: results

- SPARQL results are typically presented by tables
  - e.g. the table of patents satisfying a query
- native SPARQL commands allow
  - to select the columns in the results table
  - to order rows in the results table
  - to extract summary values from rows,
    - hence to extract statistics
- it is also possible to make inferences from your data,
  - in any case under your control

SPARQL: a simplified example

- the question to solve: identify all patents citing patent X, directly or indirectly (i.e. through a chain of citations)
- the kernel of the SPARQL query is

```
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”, which includes in this case the recursion symbol +
  - the ?s are the unknown subjects to identify
- the SELECT DISTINCT clause produces the output

Linked Data accesses

Linked Data are accessed at specific web addresses (end-point). *Depending from the end-point*, they can be:

- **linked public data**, used for
  - open and reliable vocabularies and ontologies
  - open and reliable information, e.g. DBpedia

  provided that the performance is adequate for you
- **linked restricted data**
  - data restricted to you and to your business partners
- **linked internal data** (behind your firewall) to access
  - your internal company data

Whilst Linked Data emerged first as Linked Open Data, it is misleading to assume that Linked Data have to be necessarily open as well.

intellisemantic

What is available in patents

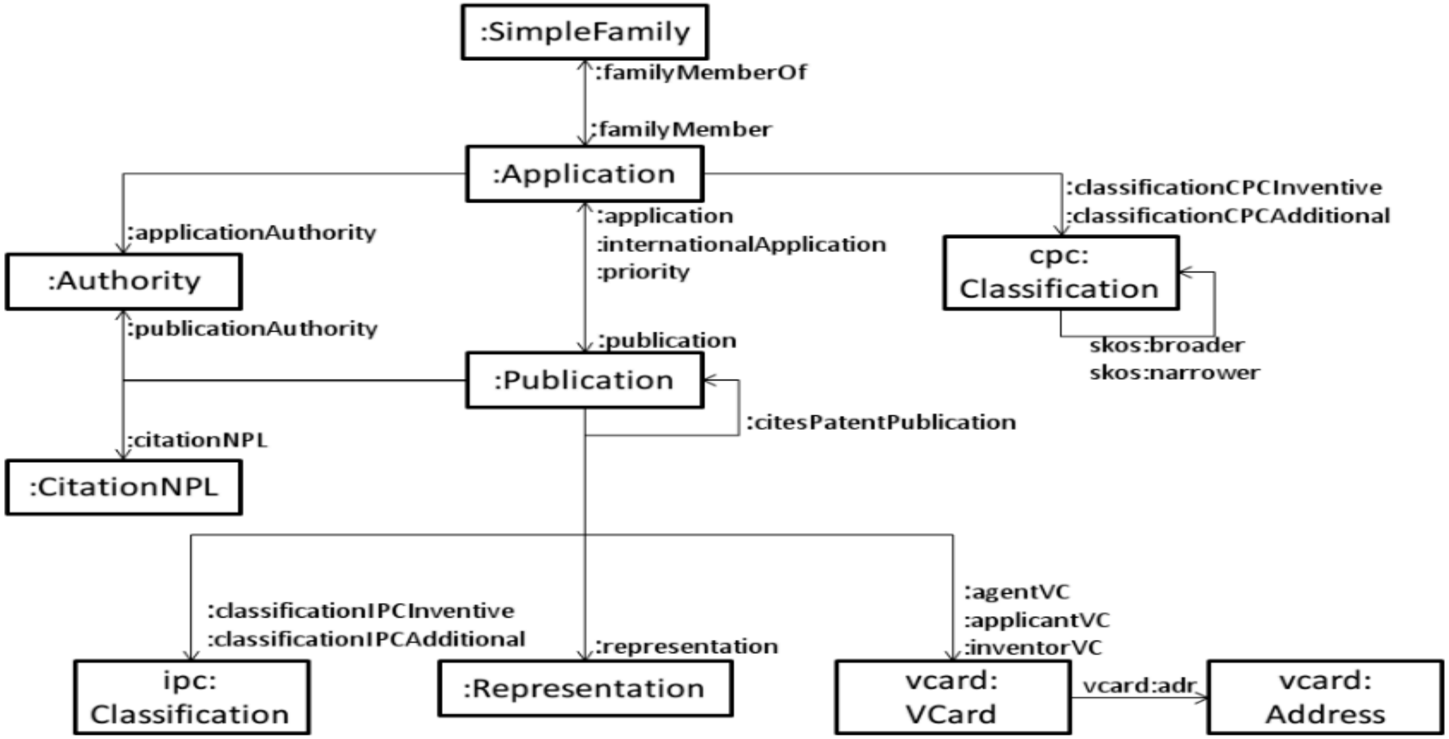
The EPO LOD: an overview

- It is one of the three kind of services provided by EPO to application integrators
- It is available since April 2018, at https://www.epo.org/searching-for-patents/data/linked-open-data.html#tab-1
- It is provided under the Creative Common Attribution 4.0 License, hence the integrator:
  - is free to share and adapt these data
  - giving credit and without adding restrictions
- User manual and support forum available
- Panel access at https://data.epo.org/linked-data/
  - it provides a SPARQL query suggestion interface
  - it provides access to a SPARQL endpoint

The EPO LOD: technical details

- The data base is **EPO-specific** and includes:
  - data provided by EPO, i.e.
    - EPO application and publication bibliographic information, **including text**
      - weekly updated
    - the CPCs tree
    - a **vocabulary of patent concepts defined by EPO**
  - links to other LOD key vocabularies/concepts
    - most of which originated by W3C
- Text search also available
  - **which is an interesting extension**
- The data base does not include (yet?):
  - information about the legal status

An example: the EPO LOD patent vocabulary



This vocabulary developed by the EPO defines some key concepts in patents (e.g. application, publication) and can be shared with other patent information provided as **linked data**. *This is only a simplified sketch, indeed.*

The EPO LOD: the business model

- free data and access
- to be used according to the fair use policy, hence for casual users and experiments:
  - *best effort policy* is applied for results:1 minute after the query, results are terminated in any case
- for production level applications, the LOD data base has to be transferred to the **server of the application integrator**
  - to the purists of the open web, this is a limit
  - for the real business, this is the example of a **realistic compromise**, which can enable some more flexibility on the business model

Other LOD examples for patents

- KIPO LOD
  - info and data at http://lod.kipo.kr/
  - SPARQL query access at http://lod.kipo.kr/data/sparql
  - the data base includes bibliographic and **legal** KIPO info
  - LOD support is mentioned the annual KIPO report
  - *it is a beta now*
- US LOD
  - info and data at https://old.datahub.io/dataset/linked-uspto-patent-data
  - SPARQL query access at http://us.patents.aksw.org/sparql
  - *it seems still a demo, since the data content is not updated recently, and the kind of licence is not mentioned*

Opportunities and challenges Conclusions

Bulk data base providers (patent offices)

- Opportunities
  - technical: the adoption of the LD paradigm facilitates the verification and cleaning of your own data base, hence it should be promoted also as an **internal best practice**
- Challenges
  - business: if a provider delivers LD to its customers, it has to define its license and its commercial conditions
    - beware: the provider is not forced by the technology itself to offer its data for free! This a misleading objection.

Application developers

- Technical opportunities:
  - the **integration of patent data bases** provided by different patent officies will become easier
  - the integration of patent data with other information (e.g. scientific or business) becomes easier
    - this will support the **extension of patent platforms to business intelligence platforms**
  - more advanced functions can be provided to patent professionals and their environment (i.e. colleagues and managers)
- Business challenges:
  - all the previous mentioned opportunities will materialize easier if the LD paradigm will suitably adopted by data base providers

Patent professionals

- Technical opportunities: have to opportunity to rely on:
  - more **powerful and diversified queries**, e.g. to suitable explore the citations network
  - more **advanced and flexible analytics tools**, for deeper patent sets analyses
  - more transparent and well agreed definition of concepts, to support a **new generation of semantic solutions** in patent informatics
    - I define this **semantic 3.0** in patent informatics
  - **more information rich tools**, which can help them to **support better a broader set** of users
- Challenges:
  - motivations and willingness to improve the profession

Semantics in Patent Informatics: a view

Semantics in Patent Informatics can be categorized also by generations:

- **Semantics 1.0 (implicit):** Solutions and best practices in patent informatics include some semantic concepts, but they are not mentioned under this term. For example: IPCs and CPCs are actually *ontologies*.
- **Semantics 2.0 (explicit)**: Natural language semantics is included in solutions, to add new ways of searching and/or to analyze results. The word "semantic" is now explicitly advertised as such, for very diversified solutions indeed. Some concerns arise too, e.g. about "which kind of semantic" is in place (for searching? for analyzing?).
- **Semantics 3.0 (standardized):** The inclusion of explicit, agreed and transparent ***standard*** knowledge bases will establish semantics as a true value added for the professional, not as a question mark issue.

Conclusions

- The adoption of Linked Data is a **win / win opportunity** for all players:
  - patent offices
  - application developers
  - professional users
  - their environment (colleagues and managers)
- but the **awareness** of this possible evolution in patent informatics **must sparkle now!**
- Linked Data in fact:
  - uses well established international **standards**
  - facilitates the **reuse of existing knowledge data bases** (e.g. vocabularies)

Questions

- technology-related questions?
- application-related questions?
- business model-related questions?
  - bulk data base providers?
  - application developers?
- other questions?

Acknowledgment and extensions

- since the topic is definitely challenging to be summarized in a 20 minute presentation, an open access paper with the same title and coauthored with Marco Ciaramella will extends this presentation with:
  - more details
  - patent specific examples
  - bibliography
- the paper will be accessed at https://arxiv.org/
  - in the section computers / digital libraries
- I do acknowledge Marco for the contribution to this paper as well to this presentation

Contact information

a) For specific questions and information requests

- e-mail: alberto.ciaramella@intellisemantic.com
- tel. +39 011 9550 380

b) For registering to the IntelliSemantic newsletter, including references to new IntelliSemantic papers and webinars announcements

- e-mail: newsletter@intellisemantic.com
- next issue: September 2018

c) For general information and first level overview:

- site http://www.intellisemantic.com