Building a comprehensive patent citations dataset

Latest Version

PatCit

G. de Rassenfosse
G. Cristelli
F. Gerotto
K. Higham
C. Verluise
L. Violon

EPFL
EPFL
SciencesPo
Hitotsubashi University
College of France
HEC Paris

Thanks
P. Lopez
T. Simcoe

You!
Sections

- Mission
- Motivations
- Previous work
- Methodology
- Results
- Data FAIR
Mission

Build a community driven comprehensive patent citations dataset

- Worldwide
- Front page & In-text
- All categories of NPL

→ One-stop-shop by and for the community + we take care of data management
Motivations

Embed patents in the innovation system & Improve knowledge flows measurement

● Embed patents in the innovation system
  ○ Science (bibliographical reference & genetic database)
  ○ IO/Competition (standards, product & litigation)
  ○ Open knowledge (Webpage, Wiki)
  ○ IP Institutions (search report & office action)

● Improve knowledge flows measurement
  ○ In-text bibliographical reference & patent citations → Improved proxy to knowledge flow
  ○ Include patents cited as NPL (!)
Previous work

We build on the shoulders of giants

- Patents and scientific citations
  - Marx and Fuegi (2020a,b)
  - Bryan et al (2020)

- Natural Language Processing & citations extraction/parsing
  - Galibert et al (2010)
  - Verberne et al (2018)
  - Grobid (2008-2020)
Methodology

Bridge the gap between the two fields

- Syndicate field experts
- Develop domain-specific information extraction models
- Enrich data
The solubilized enzyme of this process also shows lower activity (...)

Sulfonylacetate derivatives as described in U.S. Pat. No. 4,060,420, salts (...)


Sulfonylacetate derivatives as described in U.S. Pat. No. 4,060,420, salts (...)

Results

- **Front-page**
  - Coverage: worldwide
  - Categorize: 10 categories, f1: .86
  - Parse: already 4 category specific information extraction models
  - Consolidate: 40% bibliographical references enriched with Crossref

- **In-text**
  - Coverage: US all time
  - Extract: bibliographical references & patents, f1: .69 & .89
  - Parse: both categories
  - Consolidate: 33% bibliographical references enriched with Crossref and 75% patents enriched with PATSTAT/Claims
<table>
<thead>
<tr>
<th>Document Type</th>
<th>Categorization</th>
<th>Parsing</th>
<th>Consolidation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographical reference</td>
<td>p: .84</td>
<td>r: .92</td>
<td>f: .87</td>
</tr>
<tr>
<td>Office action</td>
<td>p: .89</td>
<td>r: .93</td>
<td>f: .90</td>
</tr>
<tr>
<td>Patent</td>
<td>p: .78</td>
<td>r: .89</td>
<td>f: .83</td>
</tr>
<tr>
<td>Search report</td>
<td>p: .96</td>
<td>r: .87</td>
<td>f: .91</td>
</tr>
<tr>
<td>Norm Standard</td>
<td>p: .87</td>
<td>r: .71</td>
<td>f: .78</td>
</tr>
<tr>
<td>Product documentation</td>
<td>p: .66</td>
<td>r: .62</td>
<td>f: .64</td>
</tr>
<tr>
<td>Webpage</td>
<td>p: .57</td>
<td>r: .50</td>
<td>f: .53</td>
</tr>
<tr>
<td>Database</td>
<td>p: .94</td>
<td>r: .88</td>
<td>f: .90</td>
</tr>
<tr>
<td>Litigation</td>
<td>p: .93</td>
<td>r: .84</td>
<td>f: .88</td>
</tr>
</tbody>
</table>

a=accuracy | p = precision | r = recall | f = f1-score
<table>
<thead>
<tr>
<th></th>
<th>Extraction</th>
<th>Parsing</th>
<th>Consolidation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographical reference</td>
<td>p: .80</td>
<td>r: .60</td>
<td>f: .69 [ex-post]</td>
</tr>
<tr>
<td>Patent</td>
<td>p: .97</td>
<td>r: .82</td>
<td>f: .89</td>
</tr>
</tbody>
</table>

p = precision | r = recall | f = f1-score
Ex-post: we apply a classifier on top of Grobid extraction model to improve the precision. Ex ante p: .41, r: .64, f: .51
Data FAIR

- **Find**
  - Zenodo [dump + data versioning]
  - BigQuery [interactive + latest]

- **Access**
  - Documentation website
  - Issues/requests/etc on GitHub

- **Interoperability**
  - [bibref] DOI, PMID, PMCID
  - [patent] publication number
  - [database] accession number
  - [wiki] url

- **Reproduce**
  - GitHub
  - DVC (model + training data)
Updates

Stay informed & Get involved

PatCit in your mails

v0.3 Coming soon!

PatCit on GitHub

⭐ Star  ⚒️ Fork
Thank You!
<table>
<thead>
<tr>
<th>Category</th>
<th>Group</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographical</td>
<td>Science</td>
<td>B. Katz et al., 8086 Microcomputer Bridges the Gap Between 8 and 16 Bit</td>
</tr>
<tr>
<td>Database</td>
<td>Science</td>
<td>DATABASE GENBANK [online] XP003015033, accession no. EMBL Database</td>
</tr>
<tr>
<td></td>
<td></td>
<td>accession no. (AAU25674)</td>
</tr>
<tr>
<td>Norm Standard</td>
<td>IO/Competition</td>
<td>Qualcomm Europe, sounding Reference Signals, GPP TSG RAN1 #49bis</td>
</tr>
<tr>
<td>Product documentation</td>
<td>IO/Competition</td>
<td>Brochure, Roadtec SP 100 Gravity Feed Paver (undated).</td>
</tr>
<tr>
<td>Litigation</td>
<td>IO/Competition</td>
<td>Suggestion for Interference filed in connection with U.S. Appl. No.</td>
</tr>
<tr>
<td>Category</td>
<td>Group</td>
<td>Example</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Wiki</td>
<td>Open knowledge</td>
<td>Operating System searched from Wikipedia on Oct. 8, 2010.</td>
</tr>
</tbody>
</table>
Technical stack

- PATSTAT
- Google Patents
- Grobid
  - biblio-glutton
- spaCy
- prodigy
- Google Big Query
- zenodo

Input: PatCit
Output:
Named Entity Recognition

TODO
Vocabulary

**Accuracy** - Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: \( \text{Accuracy} = \frac{\# \text{ correct predictions}}{\# \text{ of predictions}} \).

**Precision** - Informally, precision attempts to answer the following question: "What proportion of positive identifications was actually correct?". Formally, precision has the following definition: \( \text{Precision} = \frac{\# \text{ True Positives}}{\# \text{ True Positives} + \# \text{ False Positives}} \).

**Recall** - Informally, recall attempts to answer the following question: “What proportion of actual positives was identified correctly?”. Formally, recall has the following definition: \( \text{Recall} = \frac{\# \text{ True Positives}}{\# \text{ True Positives} + \# \text{ False Negatives}} \).
Paradoxically, the USPTO full-text data implicitly cover a larger share of EP families than the EP bulk dataset.
Stay informed & get involved!

**Freshest**

- Stars
- Kudos! + receive updates in your GitHub feed

- Watchers
- Receive updates in your GitHub inbox

**Release**

- Receive updates in your mail-box

**Doc**

- Documentation updates on the project website