Sample queries and tips

for PATSTAT

- EPO Worldwide Patent Statistical Database
- EPO Worldwide Legal Status Database for PATSTAT

Author: EPO - Electronic Publication and Dissemination

Doc Version 2.1

Save Date: 18 March 2016

1. General

The purpose of this document is to help you starting quickly with PATSTAT, even if you are not very familiar with PATSTAT data, SQL or the patent domain.

This document comprises several sections:

- **Sample queries** which can be executed as is. You may take them as a starting point and adapt them to your needs.
- **Tips and tricks** will help you to make the best use of PATSTAT and to avoid common pitfalls.
- **Useful resources** on PATSTAT Online, PATSTAT data and SQL are in the final section.

Your questions and comments are very welcome. Please send them to the helpdesk patentinformation@epo.org .

2. Sample queries

This section will help you to quickly get some results. You can cut & paste every query to the query field of the Search Window of PATSTAT Online or into the query editor of your own database client. You can also easily adapt the queries to your needs.

In case you run the queries on a database you created yourself with PATSTAT Raw Data, it might be necessary to modify the queries if you are not using T-SQL (MS SQL Server), but another dialect of SQL (ORACLE, Postgres, DB2, …).

For your rough estimation, the runtime of each query is classified. The run time on PATSTAT Online may vary depending on server usage, but typically it is

 less than 2 seconds
 less than 20 seconds
 more than 20 seconds

a) What are the 10 most cited applications filed in Great Britain?

```
SELECT Top 10 nb_citing_docdb_fam, appln_id, appln_auth, appln_nr,
        appln_kind, appln_filing_date
FROM tls201_appln
WHERE appln_auth = 'GB'
ORDER BY nb_citing_docdb_fam DESC
```

On 2015 Autumn data edition: 10 rows retrieved; runtime: 

This query makes use of the attribute `nb_citing_docdb_fam`, which contains the number of distinct DOCDB families citing the application or any of its DOCDB family members. The citation frequency on family level is for many purposes more significant than the number of citations on publication level.

b) Who are the most active applicants in Austria?

```
SELECT Top 10 COUNT(*) AS NumberOfApplications, doc_std_name,
        person_ctry_code
FROM tls206_person p
JOIN tls207_pers_appln pa ON p.person_id = pa.person_id
JOIN tls201_appln a ON pa.appln_id = a.appln_id
WHERE p.person_ctry_code = 'AT'
        AND pa.applt_seq_nr > 0
GROUP BY doc_std_name, person_ctry_code
ORDER BY  NumberOfApplications DESC
```

On 2015 Autumn data edition: 10 rows retrieved; runtime: 

To limit the result to applicants and to exclude persons which are inventors only, the attribute `applt_seq_nr` must be larger than 0.

Here the DOCDB standardized person names are used (attribute `doc_std_name`). You could also use other standardized names (EEE-PPAT or the OECD Harmonized Applicant Name) which are available in PATSTAT.

Note that here the applicants are restricted by their country of residence. Multi-national corporations which file centrally might bias the result.

c) Who are the Belgian applicants (wherever they file) which cooperate with applicants of another country?

```
SELECT DISTINCT p1.doc_std_name
FROM tls206_person p1
JOIN tls207_pers_appln pa1 ON p1.person_id = pa1.person_id
JOIN tls207_pers_appln pa2 ON pa1.appln_id = pa2.appln_id
JOIN tls206_person p2 ON pa2.person_id = p2.person_id
WHERE p1.person_ctry_code = 'BE'
    AND pa1.appln_id > 0
    AND pa2.appln_id > 0
    AND p1.person_ctry_code <> p2.person_ctry_code
ORDER by p1.doc_std_name
```

On 2015 Autumn data edition: 43.453 rows retrieved; runtime: 

Below is a more elaborate version. The international co-applicants are also returned. The pair of Belgian and international co-applicants are ranked according to the number of applications they filed together.

```
SELECT COUNT(*) AS numberOfCommonApplications,
    p1.doc_std_name as name1, p1.person_ctry_code as cc1,
    p2.doc_std_name as name2, p2.person_ctry_code  as cc2
FROM tls206_person p1
JOIN tls207_pers_appln pa1 ON p1.person_id = pa1.person_id
JOIN tls207_pers_appln pa2 ON pa1.appln_id = pa2.appln_id
JOIN tls206_person p2 ON pa2.person_id = p2.person_id
WHERE p1.person_ctry_code = 'BE'
    AND pa1.appln_id > 0
    AND pa2.appln_id > 0
    AND p1.person_ctry_code <> p2.person_ctry_code
GROUP by p1.doc_std_name, p1.person_ctry_code, p2.doc_std_name,
    p2.person_ctry_code
ORDER BY numberOfCommonApplications DESC, p1.doc_std_name ASC,
    p2.doc_std_name ASC
```

On 2015 Autumn data edition: 226 838 rows retrieved; runtime: 

d) Which applications were filed in Portugal where the inventor is also the applicant?

```
SELECT a.appln_id, appln_auth, appln_nr,
          appln_kind, appln_filing_date
FROM tls201_appln a
JOIN tls207_pers_appln pa ON a.appln_id = pa.appln_id
WHERE appln_auth = 'PT'
    AND (applt_seq_nr >0)
    AND (invt_seq_nr > 0)
```

On 2015 Autumn data edition: 7 205 rows retrieved; runtime: ⌛ - ⌛⌛

The term `(applt_seq_nr > 0) AND (invt_seq_nr > 0)` selects all persons which are applicant as well as inventor.

e) Counting the first filings of the company *Clinic de Barcelona* in 2009

Note: This example is based on the PATSTAT discussion forum entry
http://forums.epo.org/epo-worldwide-patent-statistical-database/topic2062.html

Of course, there may be several variations of the company's name spelling which are taken into account by using the wildcard character "%"

```
SELECT person_name, a.appln_id
FROM tls201_appln a
JOIN tls207_pers_appln pa ON a.appln_id = pa.appln_id
JOIN tls206_person p ON pa.person_id = p.person_id
WHERE a.appln_filing_date >= '2009-01-01'
    AND a.appln_filing_date <= '2009-12-31'
    AND a.appln_id = a.earliest_filing_id          -- limit to first filings
    AND pa.applt_seq_nr > 0                                -- limit to applicants
    AND p.person_name like '%clinic%barcelona%'
    ORDER by person_name
```

On 2015 Autumn data edition: 3 rows retrieved; runtime: ⌛⌛

f) Get all A1 publications published by the USPTO within Q1/2009

```
SELECT publn_auth, publn_nr, publn_kind, publn_date
FROM tls211_pat_publn
WHERE publn_auth = 'US'
    AND publn_kind = 'A1'
    AND publn_date >= '2009-01-01'
    AND publn_date <= '2009-03-31'
```

```
ORDER BY publn_date
```

On 2015 Autumn data edition: 83.700 rows retrieved; runtime: ⌛

g) Get all applications which are classified by both the IPC-symbols 'C01B' and 'H01M 4/583'

```
SELECT appln_id, appln_auth, appln_nr,
          appln_kind, appln_filing_date
FROM tls201_appln a
WHERE
   EXISTS
      (SELECT i.appln_id
      FROM  tls209_appln_ipc i
      WHERE i.appln_id = a.appln_id
      AND ipc_class_symbol LIKE 'C01B%')
   AND EXISTS
       (SELECT i.appln_id
      FROM  tls209_appln_ipc i
      WHERE i.appln_id = a.appln_id
      AND ipc_class_symbol = 'H01M   4/583')
```

On 2015 Autumn data edition: 1 131 rows retrieved; runtime: ⌛

Note the 3 spaces in the symbol `'H01M  4/583'`. The IPC (or CPC) main group (here: 4) always needs 4 positions. Then main group number is always right aligned, and the appropriate number of spaces is used to fill up4 position. This corresponds to WIPO standard ST.8.

Retrieving all applications which contain (among others) a specific IPC class / group is much easier and faster.

```
SELECT appln_id
FROM tls209_appln_ipc
WHERE ipc_class_symbol LIKE 'B60K%'
```

On 2015 Autumn data edition: 546 002 rows retrieved; runtime: ⌛

h) Which office published the most applications (filed in 2009) within 15 month?

Normally, the first publication takes place after 18 month. Here we are retrieving applications which are published significantly earlier.

```
SELECT COUNT(*) AS number, appln_auth
FROM tls201_appln
WHERE appln_filing_year = 2009
        AND dateadd(month, 15, appln_filing_date) >= earliest_publn_date
GROUP BY appln_auth
ORDER BY number DESC
```

On 2015 Autumn data edition: 91 rows retrieved; runtime: ⌛

i) Who are the inventors of the "Mayo Clinic" and in which areas are their patents which have been filed first since 2000?

```
SELECT DISTINCT person_name, person_ctry_code, person_address,
    ipc_class_symbol
FROM tls206_person p
JOIN tls207_pers_appln pa ON p.person_id = pa.person_id
JOIN tls209_appln_ipc i ON pa.appln_id = i.appln_id
WHERE pa.invt_seq_nr > 0  -- return inventors only
AND pa.appln_id IN
    (SELECT a2.appln_id
    FROM tls201_appln a2
    JOIN tls207_pers_appln pa2 ON a2.appln_id = pa2.appln_id
    JOIN tls206_person p2 ON pa2.person_id = p2.person_id
    WHERE p2.person_name LIKE '%mayo clinic%'
    AND pa2.applt_seq_nr > 0
    AND a2.earliest_filing_year >= 2000
    AND a2.earliest_filing_year < 9999)  -- to exclude invalid dates
ORDER BY p.person_name, i.ipc_class_symbol
```

On 2015 Autumn data edition: 74 rows retrieved; runtime: ⧖⧖

Here the earliest filing date (see the Data Catalog for its exact meaning) is used because this date is closer to the date on invention than the filing date.

j) Retrieve all applications of the largest patent family.

Here we are retrieving the DOCDB family (also called "simple family").

```
SELECT  docdb_family_size, docdb_family_id, appln_id, appln_auth, appln_nr,
          appln_kind, appln_filing_date
FROM tls201_appln
WHERE docdb_family_size =
    (SELECT max(docdb_family_size) FROM tls201_appln)
ORDER BY docdb_family_id, appln_filing_date, appln_auth
```

On 2015 Autumn data edition: 470 rows retrieved; runtime: ⧖

The same result can be computed without the attributes `docdb_family_size` in table `tls201_appln`, albeit the query would take longer to run:

```
SELECT  docdb_family_id, appln_id, appln_auth, appln_nr,
          appln_kind, appln_filing_date
FROM tls201_appln a
WHERE docdb_family_id =
    (SELECT TOP 1 docdb_family_id  -- here the (single) largest family is computed
    FROM tls201_appln
    WHERE DOCDB_FAMILY_ID > 0 -- exclude the dummy family
    GROUP BY docdb_family_id
    ORDER BY COUNT(*) DESC)
```

```
ORDER BY docdb_family_id, appln_filing_date, appln_auth
```

On 2015 Autumn data edition: 470 rows retrieved; runtime: ⧖⧖

Alternatively, you could use the INPADOC family, which is broader than the DOCDB family, by using attribute `inpadoc_family_id` instead of attribute `docdb_family_id`.

k) Get applications which contain both the words "bicycle" and "plastic" in title or in the abstract.

```
SELECT a.appln_id, appln_auth, appln_nr,
          appln_kind, appln_filing_date
FROM tls201_appln a
LEFT OUTER JOIN tls202_appln_title t ON a.appln_id = t.appln_id
LEFT OUTER JOIN tls203_appln_abstr abstr ON a.appln_id = abstr.appln_id
WHERE (t.appln_title LIKE '%bicycle%' AND t. appln_title LIKE '%plastic%')
    OR (abstr. appln_abstract LIKE '%bicycle%'
        AND abstr. appln_abstract LIKE '%plastic%')
```

On 2015 Autumn data edition: 2 407 rows retrieved;
runtime: ⧖⧖⧖ (about 20 minutes)

Here an OUTER JOIN is used to also retrieve applications where the search words occur in the title, but there is not abstract for this application; or the search words occur in the abstract, but the title is missing.

Please note that PATSTAT is designed to handle structured data very well, but as a trade off cannot search efficiently within texts (like title and abstract). If you need to do text search often, please consider other EPO products, like Espacenet or GPI (Global Patent Index).

3. Tips and tricks

This section helps you to better understand of the PATSTAT database structure and the patent domain. You will also learn to avoid common pitfalls.

3.1. Questions about PATSTAT data and tips for querying

- **Where can I find the detailed PATSTAT data model description?**
  The "Data Catalog" of the newest PATSTAT version can be found in the download box of www.epo.org/searching-for-patents/business/patstat.html

- **What is the data coverage of PATSTAT?**
  Generally spoken, PATSTAT is based on DOCDB: what is not in DOCDB will not be available in PATSTAT. Keep in mind the (minor) exceptions with regards to replenished applications that have been created to compensate for un-linkable (unknown) applications or publications. Also extra address information has been added from the EPO register and the USPTO register.

  General coverage information can be found in the document "Contents and coverage of the DOCDB bibliographic file". Coverage of the legal status (table `tls221_inpadoc_prs`) is described in the document "Contents and coverage of the INPADOC legal status file". Both documents can be downloaded from http://www.epo.org/searching-for-patents/helpful-resources/raw-data/data/tables/weekly.html

  Note that DOCDB is continuously updated, while PATSTAT is a "snapshot" taken about 12 weeks before production of the PATSTAT release.

- **How up-to-date is PATSTAT's data?**
  The spring version contains a snapshot of EPO's Master Bibliographic Database as of early January, the autumn version is based on the data of early August. However, typically applications are published 18 month after filing and must be kept secret before publication.

  Also, it might take some time till patent offices deliver application data and EPO processes these data, so you should include a safety margin of at least 6 month, better 18 month.

  As an example, let's assume you are working with the **spring 2016** version of PATSTAT data:
  - data snapshot - **early 2016**
  - filed about 18 month earlier - **mid 2014**
  - depending on your safety margin, you can assume PATSTAT to contain applications up to the **end of 2012**, probably **end of 2013**

- **What are artificial applications / publications?**
  Artificial applications have been created to compensate for un-linkable (unknown) applications or publications. They have an `appln_id` >= 900.000.000. For details see section "Application Replenishment" in the PATSTAT Data

Catalog.
Similarly, there are **artificial publications**, which are explained in section "Publication Replenishment" of the PATSTAT Data Catalog.

- **Do the IDs within PATSTAT change from one edition to the next?**
  IDs are introduced for technical reasons and do not convey any business meaning. The `appln_id`, `pat_publn_id`, `person_id` and some IDs more are stable, i.e. they do not change from edition to the next. Other IDs are not stable. For details see section "Surrogate Database Keys" of the Data Catalog.

- **What does date '9999-12-31' mean?**
  If for some reasons a date. e. g. a publication date, is not known, the dummy value '9999-12-31' will be assigned. If you want to retrieve data which are newer than a certain date, make sure to exclude this dummy date. For example, to retrieve all Danish publications since 2005, use a query like this:

```
SELECT *
FROM tls211_pat_publn
WHERE Publn_auth = 'DK'
    AND publn_date >= '2005-01-01'
    AND publn_date <   '9999-12-31'
```

- **There are so many dates. Which should I use?**
  This depends on your needs, but there is a rule of thumb.

  Take the *earliest filing date* if you analyse inventions, because this date is the closest to the act of invention.

  Take the *date of filing* if you are interested in application filings.

  Take the *date of the first publication* if the legal impact is of importance to your research. Note that the publication of the application already has some legal consequences.

  In any case, remember that none of these dates will be available before the first publication.

- **Should I count publications, applications or families?**
  Again, this depends on your analysis. You should be aware that there is fundamental difference whether you are counting documents (i. e. publications), applications (i. e. filings in various offices) or inventions (i. e. families, which contain groups of similar patents).

- **How can I retrieve applicants?**
  Applicants as well as inventors are stored in table `tls206_person`. This person table does not only contain physical persons, but also corporations and other organisations.

  To combine this person table (e. g. names of persons) with the core data of an

application in table `tls201_appln`, you have to join these 2 tables via the table `tls207_pers_appln`.

If you want to analyse how applicants and inventors change from one publication to the next, you must join the publication and person tables via the table `tls227_pers_publn`.

Applicants are persons whose value of the attribute `applt_seq_nr` is larger than 0, so make sure to add the condition '`applt_seq_nr > 0`' to your query. Likewise, to select inventors only use '`invt_seq_nr > 0`'.

- **How should I handle the many name variations of applicants and inventors?** The data contained in PATSTAT has been delivered from many national sources over a long time period. Because no internationally agreed unique identifier for applicants / inventors exists, very often there are several name variations for a given company, organization or individual. Several organizations tackle this problem by harmonizing names. PATSTAT offers several of these harmonized names (DOCDB standardized name, EEE-PPAT, OECD HAN).

- **How can I identify PCT applications? Which office was the Receiving Office?**

  Filings of International applications (PCT applications) at a Receiving Office can be identified in table `tls201_appln` by having `appln_kind = 'W '`. The attribute `appln_auth` denotes the Receiving Office.
  "WO" (WIPO) is generally not used as an application authority for PCT applications.

  Publications of the international application by WIPO can be identified in table `tls211_pat_publn` by having `publn_auth = 'WO'`. The `publn_nr` will contain the WO number in DOCDB format.

  International applications in the **national/regional phase** can be identified by having `internat_appln_id > 0`. In fact, the `internat_appln_id` is the same number as the `appln_id` of this application filed in the Receiving Office.

  For the 2016 Spring Edition or later editions you may simply use the attribute `int_phase` in table `tls201_appln` , which is an indicator whether an application is or has been in the international phase.

- **How can I identify EP patents which are in the national phase?**

  In most cases EP patent which entered the national phase are not re-published by the national offices. Notable exceptions are AT, DE, ES and GR.

  As a consequence, for most EP member states no publications exist for EP-patents in their national phase. Therefore these patents are not recorded in DOCDB, the main data source of PATSTAT, and consequently are not available in PATSTAT. Still, by checking the Legal Events (INPADOC PRS) in table `tls221_inpadoc_prs`, you can identify these documents by looking for the legal

status code 'PGFP' (Post Grant Fee Paid) of the EP patent. Note that his table has to be bought separately (see product 14.24.1 in http://www.epo.org/searching-for-patents/business/patstat.html)

- **Why do the application numbers look different from Espacenet?**
  The application number (`appln_nr`) in PATSTAT is in DOCDB format, because DOCDB is PATSTAT's main data source. On the other hand, the popular search application Espacenet is displaying application numbers in another format (EPODOC format).
  Luckily, the PATSTAT attribute `appln_nr_epodoc` in table `tls201_appln` holds the application number in EPODIC format, which you can copy & paste as-is into the Espacenet search application.

- **Some data seems to be missing. Why? What should I do?**
  PATSTAT is primarily based on DOCDB data. DOCDB data comes from different national and regional offices, which provide data in various qualities and degrees of completeness.

  As an arbitrary example, the country of residence is missing for almost all applicants / inventors of applications filed in JP.

  Therefore before starting any major analysis we recommend you to test the data you need. One possible way to overcome such problems is to take missing data from family members.

3.2. Questions about the patent domain

- **What are patent families? What are the differences between a simple/DOCDB family and an extended/INPADOC family?**
  A good explanation can be found in http://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families.html

- **What are IPC and CPC?**
  For the general introduction into patent classification see http://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification.html .

  The **IPC** (International Patent Classification) is a hierarchical system of symbols which is globally used to classify patents and utility models according to their technological area (http://www.wipo.int/classifications/ipc/en/).

  The Cooperative Patent Classification **CPC** is the outcome of an ambitious harmonization effort to bring the best practices from each EPO and USPTO together. It harmonizes the former USPTO classification and EPO's ECLA and has been introduced in Jan 2013. CPC is compatible with IPC, but more detailed. Information on CPC can be found in www.cooperativepatentclassification.org

- **Kind codes, publication codes, legal status codes are all patent office specific. Where can I find an overview?**
  This link might help you: http://www.epo.org/searching-for-patents/helpful-resources/raw-data/data/tables/regular.html

- **Do I have to consider national differences?**
  Unfortunately, yes. Although you might be familiar with the "big lines" when it comes to procedural steps for EP, WIPO or US applications, each country or organisation has its own particularities, which also may change over time. National legislation and consequently applicant behaviour can skew statistics and figures in sometimes unexpected ways. Here are just some issues that you may have to consider:
  ▪ Unity of invention
  ▪ Dual filings
  ▪ Technical relations
  ▪ Re-publications of granted regional patents
  ▪ Deferred Examination
  ▪ Professors Privilege
  ▪ …

  Example:
  Usually applications filed in JP have fewer claims than applications filed elsewhere. On average, one application filed at the USPTO is broken down into 3 applications when filed in JP. In this aspect, KR is similar to JP.

  Example:
  Due to legal changes, there is significant increase in publications of applications after the year 2000 in the USPTO.

  There are various sources to get information about the patent system of specific countries. Some free resources are:

International	
Asia, India and Saudi Arabia	EPO's Virtual Helpdesk - Asia and beyond http://www.epo.org/searching-for-patents/helpful-resources/asian.html
Individual Countries	
Malaysia	Patent Information News, Issue 3/2011, p. 10 http://www.epo.org/service-support/publications.html?id=105
Singapore	Patent Information News, Issue 4/2011, p. 10 http://www.epo.org/service-support/publications.html?id=105

4. Useful resources

PATSTAT data

- Data Catalog
  http://www.epo.org/searching-for-patents/business/patstat.html
  The authoritative source of PATSTAT's data content and structure.

- PATSTAT Data Elements
  http://www.epo.org/searching-for-patents/business/patstat.html
  An overview describing how the elements of a publication relate to PATSTAT.

PATSTAT Online

- User Manual
  http://www.epo.org/searching-for-patents/business/patstat.html
  How to use the search tool.

SQL (MS SQL) query language

Whether you are new to SQL or you are switching over from another database management system: there are numerous books and Internet sites available.

- SQL self-study course
  http://www.epo.org/searching-for-patents/business/patstat.html
  An introduction in SQL, with lots of examples ready to run on PATSTAT Online

Human support

- User support
  patentinformation@epo.org

- PATSTAT discussion forum
  forums.epo.org/patstat/

- Annual Conference "IP Statistics for Decision Makers"
  http://www.epo.org/learning-events/events/conferences.html (see also "Archive" sub-pages)

Patent Statistic

- EPO's FAQs on Patent Statistics
  http://www.epo.org/service-support/faq/searching-patents/statistics.html

- OECD Patent Statistics Manual (2009)
  www.oecd.org/sti/innovationinsciencetechnologyandindustry/oecdpatentstatisticsmanual.htm
  It addresses issues regarding the complexity of patent data and provides statisticians and analysts with guidelines for building and analysing patent-related indicators.

- Compendium of Patent Statistics (OECD, 2008)
  https://www.oecd.org/sti/inno/37569377.pdf
  Contains example of patent indicators and statistical charts.