Big Data & Public Databases for Patent Research and Analysis

Lucy Antunes, Ford Khorsandian, Ellen Krabbe, Steve Sampson, Ian Wetherbee

The growing availability of big data has made it possible to get answers in seconds to questions that previous took days of manual work. In this article we will present an overview of how big data works, what types of intellectual property-related questions and answers can be gained from analyzing it, and how to start using it today for corporations, attorneys, or other patent information professionals.

➢ What is Big Data?

Definition: (typically large) computer-readable collections of data. The term "big data" also includes the analysis of datasets to extract insights, trends, patterns and associations, even though the underlying dataset may be small.

Analyzing these datasets requires special analysis tools and computing power potentially beyond what standard relational databases are capable of. The collection of international patent metadata (dates, inventors, etc.) is typically small and can fit on a laptop, although performing complex queries of the data using a laptop may take hours or days while products designed for analyzing big data can calculate answers in seconds.

Big data can be analyzed for insights that lead to better decisions and strategic business moves.

Publicly available sources include both patent and non-patent sources. Several sources include:

EPO (OPS, PATSTAT, Linked Open Data, Espacenet RSS), USPTO (bulk data, PTAB, PEDS), ChEMBL, PubChem, World Bank, Google Patents Public Datasets, Patentscope RSS. US-centric datasets can be found via data.gov.

➢ What is required to perform big data analysis?

After identifying the potential sources of data, you will need to consider the following:

- How to acquire and store it
- How to manage and update it
- How to link it with private or public data
- How to analyze it
- How to use any insights obtained

Informational technology considerations include:

- Fast processors
- Sources for cheap, abundant, secure storage
- Cloud computing SaaS offerings
- Sharing data and compute access

➢ Why do I need a big data platform?

Big data platforms are designed to store and analyze massive amounts of information. You can run infrastructure yourself using several Apache open-source projects (Hadoop, Storm, Beam, HBase). Cloud providers (Amazon, Google, Microsoft) also sell several products for raw data storage, batch and streaming data processing, data analytics warehouses, and online databases.

For the target use case of custom patent data analysis, the data analytics warehouses are the best fit. These cloud services are designed to process ad-hoc data analysis queries over large volumes of data quickly and cheaply.

For example purposes, we have chosen Google BigQuery to demonstrate how big data can be used in answering queries about patent information. Using this platform, multiple sources of public, paid and private data may be stored and analyzed. BigQuery uses SQL, a standard, flexible language for accessing and manipulating databases. BigQuery comes with several free datasets available for anyone to query. One of these is the Google Patents Public Datasets, provided for free by IFI CLAIMS Patent Services and Google, which includes 17 countries of patent publications and is updated quarterly.

BigQuery is a cloud based account with no server setup, permitting instant querying, sharing and distribution of data. It is especially useful for those who do not have access to currently available subscription analysis tools and those interested in integrating proprietary, private and public information. Subscription data providers can offer curated data that may also be integrated into a big data platform like BigQuery to more easily distribute it to customers for analysis.

➢ What patent questions might I get answered?

Here are just a few examples.

- Corporate portfolio by year, by technology
- Patent abandon rates
- Average time from filing to disposition
- Landscape by year of assignee, inventors, key terms
- Partnering opportunities from citations or inventors
- Legal status reports
- Analysis of family members and citations
- Expiration dates of patents in a portfolio
- Percentage of patents have more than one inventor
- What funding does the government provide to promote innovation in certain patent areas?

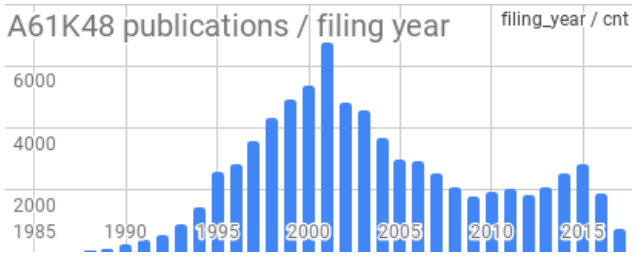
A few examples are included on the following page.

Big Data & Public Databases for Patent Research and Analysis

➢ Example 1: A61K48 publications per filing year

This is a standard query you could also perform in many patent search engines, but it is useful as a small introduction to SQL and it's capabilities.

```
#standardSQL
SELECT
COUNT(*) AS cnt,
CAST(FLOOR(filing_date / 10000) AS INT64) AS
filing_year
FROM `patents-public-data.patents.publications`
WHERE (SELECT MAX(TRUE) FROM UNNEST(cpc) AS c WHERE
REGEXP_CONTAINS(c.code, "A61K48"))
GROUP BY filing_year
ORDER BY filing_year ASC;
```

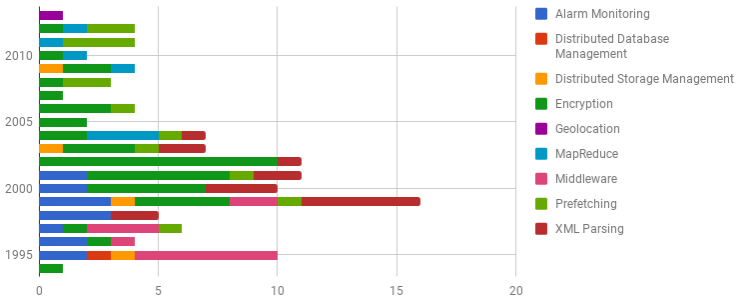


Source: "Google Patents Public Data" by IFI CLAIMS Patent Services and Google, CC BY 4.0.

While web tools are usually limited to a number of fixed graphs and dimensions of aggregation (date, assignee, inventor, CPC), with SQL you can aggregate and download any data that exists in the database. For example, you could explore the number of inventors versus the number of 102 rejections.

➢ Example 2: Portfolio filings per year

This simple query graphs the patents in your portfolio by their filing date and hand-labeled tags. As an example portfolio, this is a partial list of Google's OPN patents that was uploaded into a BigQuery table. Querying your own portfolio is as easy as matching the number format, uploading a CSV file, and setting access permissions. You can quickly export your result set to Google Sheets or download as CSV to create graphs or continue processing the data table for your reports.



Sources: "Google Patents Public Data" by IFI CLAIMS Patent Services and Google, used under CC BY 4.0, "Google OPN list" by Google.

➢ Example 3: Forward citations to a set of patents grouped by assignee and classification

One method of landscaping is to see what other assignees or inventors cite certain patents. This example aggregates the forward citations to "Standard Oil Co" patents.

Row	citing_assignee	num_cites	citing_cpc_subclass	
1	EXXONMOBIL RES & ENG CO	571	C10M	LUBRICATING COMPOSITIONS; USE OF CHEMICAL SUBSTAI
2	AFTON CHEMICAL CORP	557	C10M	LUBRICATING COMPOSITIONS; USE OF CHEMICAL SUBSTAI
3	LUBRIZOL CORP	546	C10M	LUBRICATING COMPOSITIONS; USE OF CHEMICAL SUBSTAI
4	EASTMAN CHEM CO	535	C07C	ACYCLIC OR CARBOCYCLIC COMPOUNDS
5	EXXON RESEARCH ENGINEERING CO	457	C10G	CRACKING HYDROCARBON OILS; PRODUCTION OF LIQUID
6	SHELL OIL CO	453	E21B	EARTH DRILLING, e.g. DEEP DRILLING; OBTAINING OIL, GAS
7	EASTMAN CHEM CO	405	B01J	CHEMICAL OR PHYSICAL PROCESSES, e.g. CATALYSIS, COL
8	STANDARD OIL DEV CO	353	B01J	CHEMICAL OR PHYSICAL PROCESSES, e.g. CATALYSIS, COL
9	PHILLIPS PETROLEUM CO	322	C07C	ACYCLIC OR CARBOCYCLIC COMPOUNDS
10	EXXON CHEMICAL PATENTS INC	281	C10M	LUBRICATING COMPOSITIONS; USE OF CHEMICAL SUBSTAI
11	PHILLIPS PETROLEUM CO	275	B01J	CHEMICAL OR PHYSICAL PROCESSES, e.g. CATALYSIS, COL

Sources: "Google Patents Public Data" by IFI CLAIMS Patent Services and Google, used under CC BY 4.0, "Cooperative Patent Classification" by the EPO and USPTO, for public use.

➢ Example 4: Integrating USITC (trade) data

There are sources of big data outside of patent information that can be analyzed with patent information. In this example, unfair import trade cases from the US International Trade Commission are aggregated by the WIPO technology field classification of the patents involved in the case.

Row	title	cnt
1	Computer technology	349
2	Digital communication	256
3	Audio-visual technology	251
4	Telecommunications	176
5	Semiconductors	147
6	Electrical machinery, apparatus, energy	139
7	Medical technology	104
8	Optics	100
9	Textile and paper machines	81
10	Measurement	78
11	Control	73
12	Basic communication processes	70
13	Furniture, games	67

Sources: "PatentsView" by the USPTO, US Department of Agriculture (USDA), the Center for the Science of Science and Innovation Policy, New York University, the University of California at Berkeley, Twin Arch Technologies, and Periscopic, used under CC BY 4.0, "US International Trade Commission 337 Info Unfair Import Investigations Information System" by the USITC, for public use.

Each time a new source of information is added, it can be linked with any existing dataset in the system the user can access to gain new insights.

➢ Other Resources

For an in-depth walkthrough, see the Google Patents Public Datasets blog and the BigQuery Quick Start.

WIPO Manual on Open Source Tools for Patent Analytics
50 Amazing Free Data Sources You Should Know
EPO PATSTAT