A Sloan Foundation Industry Center

The Center for Paper Business and Industry Studies

Research-Led Business-Focused

PAPERPAT: ONLINE DATABASE OF U.S. PULP AND PAPER PATENTS

A web tool for CPBIS Research

Weblink

http://www.cpbis.gatech.edu/patents

Database Description

This database is a compiling of 63,497 U.S. patents related to pulp and paper. The database covers practically all the pulp and paper patents that were issued between January 1, 1976 and December 31, 2003 by the U.S. Patent Office (USPTO).

How Was This Database Constructed?

The database was constructed using many steps.

First we established several lists of the major players (companies or organizations) in the U.S. pulp and paper industry and separated them in different categories.

The categories are:

1. U.S. pulp and paper companies in Millsonline as of 2005; major foreign companies (Canadian, Scandinavian) having subsidiaries in the U.S are not in this category
2. Old and merged U.S pulp and paper companies from Forest Product Laboratory database and from MillsOnline between 2000 and 2004
2.5 Old and merged U.S pulp and paper companies from Forest Product Laboratory database and MillsOnline between 1970 and 1999
3. Foreign Pulp and Paper Companies
4. Suppliers
5. Research and Development Organizations; universities
6. Unknown (assignee's belonging to any category above is unknown)

In the results to the query page, the categories are color-coded to facilitate identification; cf. figure 1.
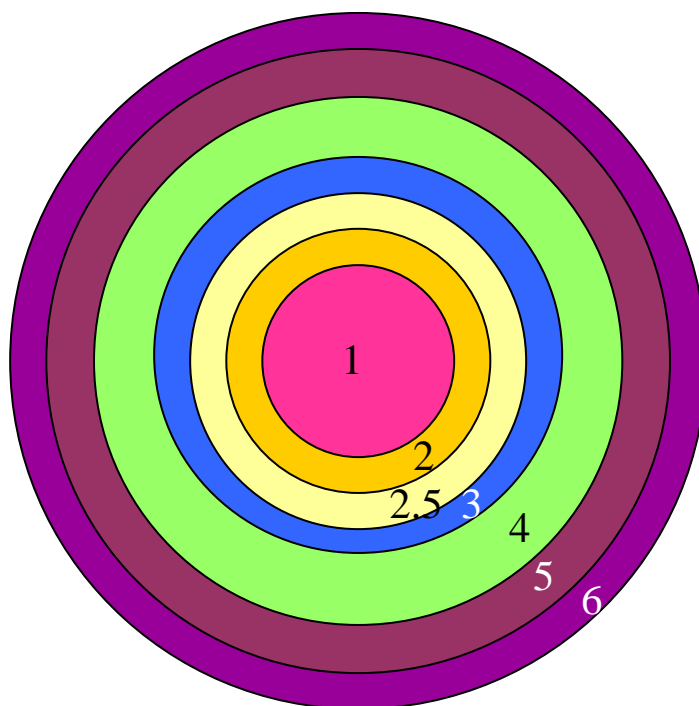
Figure 1: color coding of the different categories of patent assignees.

The company names were obtained from various sources such as MillsOnline, a Forest Products Laboratory database, paper industry websites, paper industry directories, etc.

Second, we matched as many as possible of these company names to their corresponding assignee names and assignee identification number at the U.S Patent office.

Third, an automatic search in the U.S. patent office of all the patents issued to these assignee names returned a file of over 120,000 patents. A given assignee identification number correspond to one unique assignee name but a given company can have several assignee names and identification numbers if it has separated its patent into different portfolios for instance.

We were able to restrict the number of patents in the database by using frequently used keywords. In a fourth step, we compiled a list of pulp and paper keywords using data from a pre-existing patent file.

And fifth, only those patents which contained at least one of those pulp and paper keywords in their title were kept after automatic filtering using programming. This reduced the number of "pulp and paper" patents to about 63,497 patents. Due to large role played by chemistry in the pulping and papermaking processes we estimate that non-paper-related chemical processes patents have inevitably been left even after the filtering. In the future we will be working on eliminating more non-pulp and paper patents from the pool of patents.

The following queries are available online from PaperPat;

In http://www.cpbis.gatech.edu/patents/patentskeywordsearch.php
Query by:
+ keyword(s) in the patent title (up to 4 OR/AND keywords)

or in http://www.cpbis.gatech.edu/patents/patent_exactsearch.php
Query by (a maximum of 2 queries can be combined together to restrict search):
+ full patent title or consecutive string of words in the title
+ assignee name
+ last name of author, no matter if it is the first or the n'th author (Note 1: a query with "smith", besides all patents authored by Smith, will also return patents where "Smithsonian" is an author; Note 2: for all patents authored by Gary Baum for instance, use "Baum, Gary A." as your query, to exclude other patents authored by other authors that have "Baum" as their last name.)
+ category of assignee (U.S. pulp and paper company, supplier, etc.)
+ exact patent number (without the commas; i.e. use 4046367 and not 4,046,367)

The queries will always return the results by increasing order of patent number, the lowest patent numbers (oldest patents) being at the top of the page.
All patents numbers are clickable and when clicked will open a new web page with the content of the corresponding patent on the USPTO web site, with the abstract of the patent, etc.

Source file

Specifically for use by CPBIS researchers, BoE or IAB members, or Employees of CPBIS Sponsors , a Microsoft Access file is available with significantly more information that what is available from the web tool. This is to allow our users to delve into more details into the data.
To obtain the file, please e-mail: emmanuel.lafond@cpbis.gatech.edu.

The Microsoft Access file is organized such that one row is dedicated to each patent, and the order of the rows is by increasing U.S. patent number from 1976 to 2003. The 17 columns/fields available are:

- Patent number
- 1st author
- 2nd author
- 3rd author
- 4th author
- 5th author
- 6th author (7th and higher authors aren't available)
- Assignee identification number
- Assignee name

- Category of assignee (U.S. pulp and paper maker in 2005, supplier, R&D centers, etc..)
- Examiner
- Patent application date (in Stata format; number of days since Jan 1, 1960)
- Patent issue date (in Stata format; number of days since Jan 1, 1960)
- Number of Claims
- Class(es) of the patent (according to U.S. patent classification codes)
- Title of patent
- Web link to patent page, abstract etc., on the U.S. patent office web site

Before using this data file for statistics, CPBIS recommends filtering the data file by applying filters accepting only some targeted pulp and paper keywords present in the title of the patents.

## *Data Availability Check and Accuracy Check*

The data was checked for Availability (no missing data, all entries available) and Accuracy (data is exact) and with a sampling of 1% of the database, covering 635 patents.
The results of the check are as follows, for entries:

- Patent number: available in 100 % of entries
- 1st author name: available in 100 % of entries but some rare data accuracy problems such as truncated name displayed in 2nd author name
- Assignee identification number: 5 missing, correspondiong to "withdrawn" status patents, present in 99.2% of entries
- Assignee name: 5 missing present in 99.2% of entries, missing ones not entirely connected to withdrawn patents
- Category of assignee: available in 100% of all cat.; accuracy of category of assignee:
    - cat 1: exact in 234 entries out of 243   96.3%
    - cat 2 exact in 23 entries out of 23   100%
    - cat 2.5 exact in 42 entries out of 42   100%
    - cat 3 exact in 13 entries out of 13   100%
    - cat 4 exact in 285 entries out of 285   100%
    - cat 5 exact in 26 entries out of 26   100%
    - cat 6 exact (unknow category) in 3 entries out of 3   100%

o distribution by categories of patents in sample:

Distribution of sampled patents, by categories
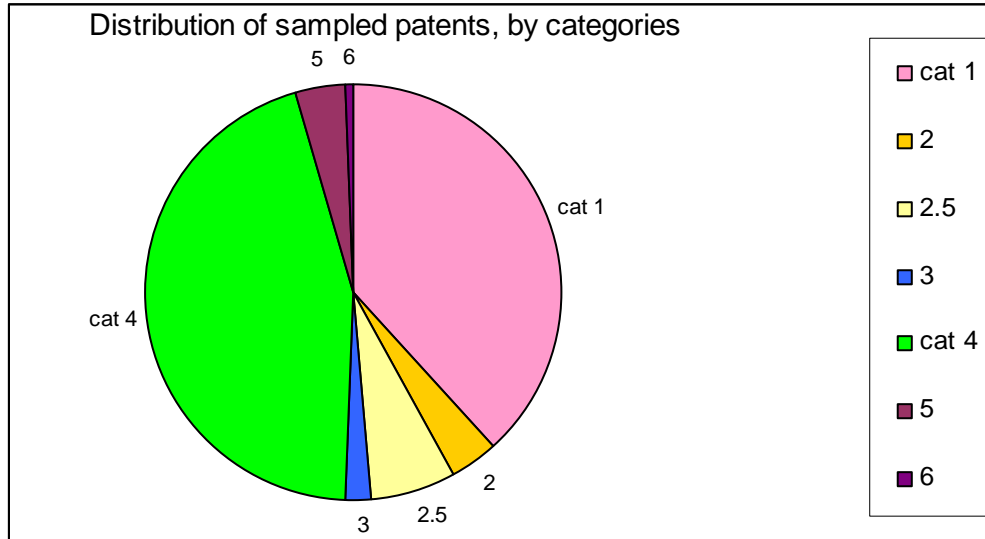
- Examiner: available in 100% of entries
- Patent application date: available in 100% of entries
- Patent issue date: available in 100% of entries
- Number of Claims: available in 100% of entries
- Class(es) of the patent: available except for 5 entries, 99.2% of entries; because of "withdrawn" status in this column
- Title of patent: available in 100% of entries;
  - Accuracy (likely a pulp and paper patent of not) tested on 159 patent titles representing 0.25% of total database: 48 (30%) are probably not pulp and paper patents and 111 (70%) are probably pulp and paper patents
  - In the "not a pulp and paper patents" categories the most represented are 1 (Eastman Kodak, Procter & Gamble for instance) and 4 (mostly chemical suppliers)
- Web link to patent page, abstract etc., on the U.S. patent office web site: available in 100% of entries

Related Work

This database is related in part to the "Role and Value of Innovation" CPBIS Research project from Prof. V. Ghosal and U. Nair-Reichert. More information on this project can be found at:
http://www.cpbis.gatech.edu/research/projects/innovation/index.htm