Patent application bibliographic data documentation

Bhaven N. Sampat
Assistant Professor
Mailman School of Public Health
Columbia University
bns3@columbia.edu

August 5, 2011

1 Summary

This file BASICBIB includes basic "front page" data for patents issued from January 1, 1975 to December 31, 2010. It is based mainly on information from a custom extract DVD generated by the Electronic Information Products Division of the USPTO on 4/29/2011. Two of the variables (application number and application date) were were extracted from the USPTO's *Cassis* BIB Patents DVD.

The dataset includes patent numbers, issue dates, application dates, application numbers, country of origin of first named inventor, first named assignee, ownership assignment category, and the primary class and subclass of each patent. It also includes four constructed variables: the number of claims, the number of forward citations to the patent (i.e. citations in subsequently issued patents), the number of backward citations to previous U.S. utility patents, and the number of citations to non-patent references.

Beyond conversation and construction of the variables indicated above, I have done little processing or cleaning of the raw data (e.g. assignee names or inventor geographic data). The NBER patent citation project

http://sites.google.com/site/patentdataproject/ is working on all of this, and much more.

For detailed listings of patent citations to and from each the patents in this dataset, use the file `USCITES` instead.

2 Codebook

```
. codebook
```

---

patent — Patent Number

---

type:	string (str7)		
unique values:	3995847	missing "":	0/3995847
examples:	"4658212"		
	"5461142"		
	"6268625"		
	"7071030"		

---

st_country — State/Country Code

---

type:	string (str3)		
unique values:	233	missing "":	0/3995847
examples:	"DEX"		
	"INX"		
	"JPX"		
	"NY"		

---

assignee — Assignee Code

---

type:	numeric (float)		
range:	[0,983310]	units:	1
unique values:	244839	missing .:	0/3995847
mean:	402017		
std. dev:	288111		

percentiles:	10%	25%	50%	75%	90%
	0	131370	395735	683245	781485

---

asstype — Patent Ownership Category at Time of Grant (Based on First Assignee)

---

type:	numeric (byte)		
label:	asstype		
range:	[1,7]	units:	1
unique values:	7	missing .:	0/3995847

tabulation:	Freq.	Numeric	Label
	5.5e+05	1	Unassigned
	1.8e+06	2	US Non-Govt Organization
	1.6e+06	3	Foreign Non-Govt Organization
	21939	4	U.S. Individual

```
                    11930         5  Foreign Individual
                    37541         6  U.S. Federal Govt
                     9399         7  Foreign Govt
```

```
class                       Primary (Original) U.S. Patent class as of 12/31/10
```

```
                type:  string (str3)
        unique values:  455                     missing "":  0/3995847
            examples:  "177"
                       "324"
                       "386"
                       "501"
```

```
subclass                Primary (Original) U.S. Patent subclass as of 12/31/10
```

```
                type:  string (str6)
        unique values:  16909                   missing "":  0/3995847
            examples:  "065000"
                       "162000"
                       "285100"
                       "462160"
```

```
issuedate                                                        Issue Date
```

```
                type:  numeric daily date (float)
               range:  [5485,18624]                 units:  1
      or equivalently:  [07jan1975,28dec2010]       units:  days
        unique values:  1878                    missing .:  0/3995847
                mean:   13486.1 = 03dec1996 (+ 3 hours)
            std. dev:   3643.94
         percentiles:        10%       25%       50%       75%       90%
                            7746     10812     14263     16496     17868
                          17mar1981 08aug1989 19jan1999 01mar2005 02dec2008
```

```
nonutility                 Patent is not a utility patent (generally reissues)
```

```
                type:  numeric (float)
               range:  [0,1]                        units:  1
        unique values:  2                       missing .:  0/3995847
           tabulation:  Freq.  Value
                       4.0e+06 0
                        12770  1
```

```
appnum                                      Series code plus serial number
```

```
                type:  numeric (long)
               range:  [108,29356432]               units:  1
        unique values:  3932708                 missing .:  7/3995847
                mean:   8.7e+06
            std. dev:   1.8e+06
         percentiles:        10%       25%       50%       75%       90%
                          6.2e+06   7.3e+06   8.7e+06   1.0e+07   1.1e+07
```

application_date — Filing Date

type:	numeric daily date (float)		
range:	[-21580,18521]	units:	1
or equivalently:	[01dec1900,16sep2010]	units:	days
unique values:	9264	missing .:	2/3995847
mean:	12608.5 = 09jul1994 (+ 11 hours)		
std. dev:	3493.46		

percentiles:	10%	25%	50%	75%	90%
	6999	10166	13459	15440	16576
	01mar1979	01nov1987	06nov1996	10apr2002	20may2005

n_claims — Number of Claims

type:	numeric (float)		
range:	[1,887]	units:	1
unique values:	369	missing .:	1065/3995847
mean:	14.8415		
std. dev:	12.6462		

percentiles:	10%	25%	50%	75%	90%
	4	7	12	20	28

claims_info_missing — Claims Information Missing

type:	numeric (float)		
range:	[0,1]	units:	1
unique values:	2	missing .:	0/3995847

tabulation:

Freq.	Value
4.0e+06	0
1065	1

n_bcites — Number of Backward Cites to Utility Patents

type:	numeric (double)		
range:	[0,1527]	units:	1
unique values:	896	missing .:	0/3995847
mean:	11.0802		
std. dev:	22.5589		

percentiles:	10%	25%	50%	75%	90%
	2	4	6	11	21

n_fcites — Number of Forward Cites as of 12/31/2010

type:	numeric (double)		
range:	[0,2214]	units:	1
unique values:	630	missing .:	12770/3995847
mean:	9.29508		
std. dev:	18.1827		

percentiles:	10%	25%	50%	75%	90%
	0	1	4	11	23

n_nplcites Number of Backward Cites to Non-Patent Literature

type:	numeric (int)				
range:	[0,2689]		units:	1	
unique values:	784		missing .:	0/3995847	
mean:	2.91181				
std. dev:	15.4297				
percentiles:	10%	25%	50%	75%	90%
	0	0	0	1	5

3 Data construction notes

This file includes name and location information on the first named inventor. The separate ALLINVENTOR file includes information for all named inventors. The patent class information in this file includes only first listed class and subclass. For a full list, see the separate ALLCLASS file. Application numbers can be merged to the separate published applications file APPLICATIONS_20012010_BIB which includes patent applications published since 2001.

Forward citation counts are based on all citing patents issued by the end of 2010, in the USCITES file. Note that these forward citation counts are not calculated, and reported as missing values, for non-utility patents. See the documentation for the citations file for details.

Backward patent citations are also constructed from USCITES. Counts of non-patent references are based on information from raw USPTO data. These raw data, and a separate file indicating patent-PMID linkages for non-patent references listed in MEDLINE, will be posted soon.

Eventually, data on the priority dates for each of the patents in this dataset will also be posted.