DCICSS/2017/2

Dynamics of Cumulative Innovation in Complex Social Systems
DCICSS PROJECT

PATSTAT1400
PATSTAT Multi-version Combined Database

Lorenzo Napolitano, Emanuele Pugliese

University of Bath
Institute for Complex System NRC (Italy), UOS Sapienza

June 12, 2017

Introduction

Our purpose is to create a unified dataset to map each patent in all the patent offices to the codes it had in different classifications *in different years*, to track not only classification changes but also the detailed flow of patents in the ever-changing classification.

Raw Data

The raw data has been provided to us by the PATSTAT officers. It involves a relevant part of the data extracted from 20 versions of PATSTAT, two versions per year (one in spring, version a, and one in autumn, version b) from 2007b to 2017a.

For each version we have tables 201 (Applications), 219 (INPADOC families), 209, 222, 223, 224, and 210 (respectively IPC, Japan, USA, CPC and other national classification codes).

Organization of data

After some preliminary analysis of the raw data, we decided to organize the data as a unified MySQL database, easy to access. We aim to have a unified structure for tables 201 and 219 along the versions (application and families) and different tables 209, 222, 223, 224 and 210 (the different classifications) for each version of PATSTAT.

First Results

This section presents the first results concerning the extent of patent reclassification across versions of PATSTAT. In this initial phase, we have focused on reclassification according the International Patent Classification (IPC) and have not yet taken other classifications schemes into account. One of the main reasons behind this decision is that the IPC is by far the most widely used technological classification. To our best knowledge, all the

national and regional patent offices considered in PATSTAT have been using it to classify patent claims for several years, which allows us to include the largest possible sample of patent applications in the tally. Note that also the USPTO, which had its own classification – the USPC – has been using the IPC since it introduction to classify all the patent applications it received. In particular, the USPTO has used the IPC and the USPS in parallel until the USPC was finally discontinued in 2013. The second advantage of the IPC is that it is has been included in PATSTAT since the first version, thus allowing to extend the analysis as far back as 2007. The only other major classification scheme that is part of every PATSTAT version since 2007 is the Japanese classification, which however only covers applications filed at the JPO.

It might be interesting to study reclassification according to other classifications in the future. This will be possible once PATSTAT1400 is completely set-up. The best candidate to replicate the exercise is probably the Cooperative Patent Classification (CPC), which has been created with the aim to create a standard worldwide classification scheme harmonizing industry- and profession-oriented schemes, like the IPC, and structure- and function-based classifications, such as the USPC [3]. The main drawback associated with the CPC is that was first introduced in the 2013 spring edition of PATSTAT, so that the available time window for the analysis of reclassification is much shorter for the CPC than the one allowed by the IPC. Extending the analysis the the USPC, which is the oldest and perhaps the most frequently employed classification in academic papers concerned using technological, does not instead seem a convenient option, since, contrary to the IPC, it does not have a clear nested structure [3]. This would would make it particularly hard, once the resolution at which to study reclassification is identified, to establish how to group codes appropriately. The example employed by [3] helps clarify this point. Class 435 in the USPC, which is used to classify patents in the field of 'chemistry: molecular biology and microbiology', contains many *sub-codes* (*e.g.* 435/35) defining a hierarchy wihin the field identified by class 435. The problem with the USPC lies in the fact that, for example, code "435/40 is a subset of 435/39, which in turn is a subclass of 435/34, but 435/41 is not a subclass of any of these". This is a big difference with respect to the IPC classification, where chemistry-related codes C12M/12, C12N/14, and C12N/60 identify three different groups, which fall into two subclasses – C12M and C12N – and one class – C12 [3]. This implies that the IPC (like the CPC) leaves the researcher the freedom to explore reclassification at many level, whereas the USPC virtually constrains the analysis in this respect.

In order to obtain a first measurement of the relevance of re-classification across PATSTAT versions, we count the number of unique (application, IPC-code) pairs that are added and dropped between consecutive versions. Furthermore, we count how many applications are interested by reclassification events from one version to the next to have another point of view on the importance of the phenomenon. For the sake of comparability, we include in the count only patent applications that are present in all the available PATSTAT versions, from version 2007b to version 2017a. Of course, this implies that we are excluding from the exercise all patents added to the database after 2007. As shown in figure 1, this choice has a considerable cost in terms of sample size, since the number of patent applications has grown almost steadily over time. However, we are still left with with a large sample of over 45 million patent applications associated to at least one IPC code. The plots in figure 1 also show two unexpected, and possibly interesting features, which deserve future investigation. In particular, there is a very noticeable dip in the number of (application, IPC code) pairs (dashed line) between version 2010b and version 2011a. What makes the

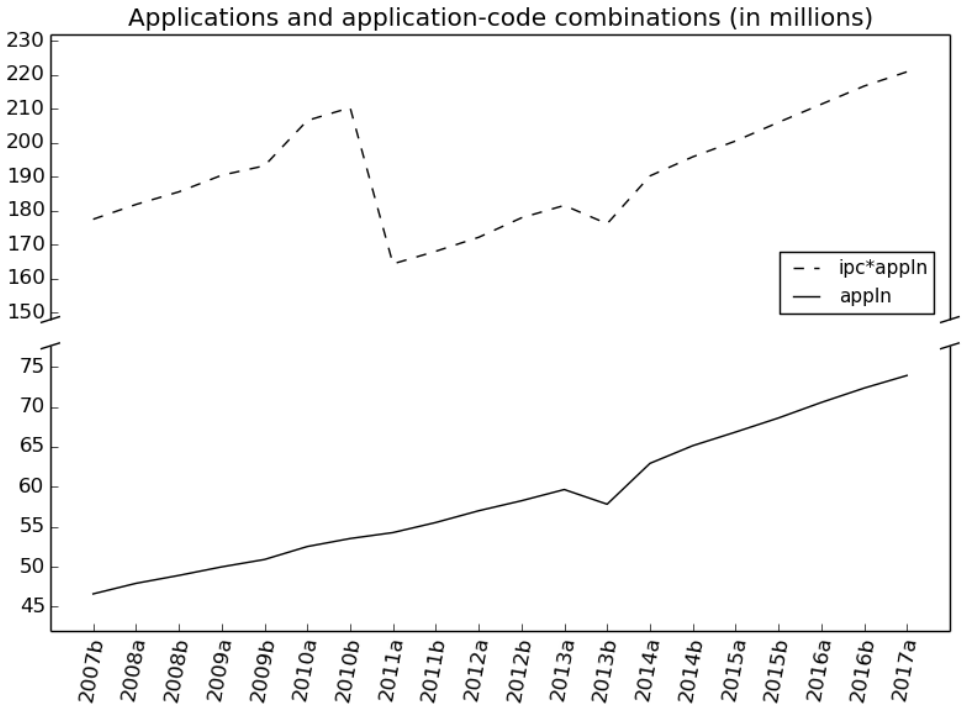


Figure 1: Number of applications with at least one claim (bottom) and number of unique (application, IPC subgroup) pairs across PATSTAT versions (top)

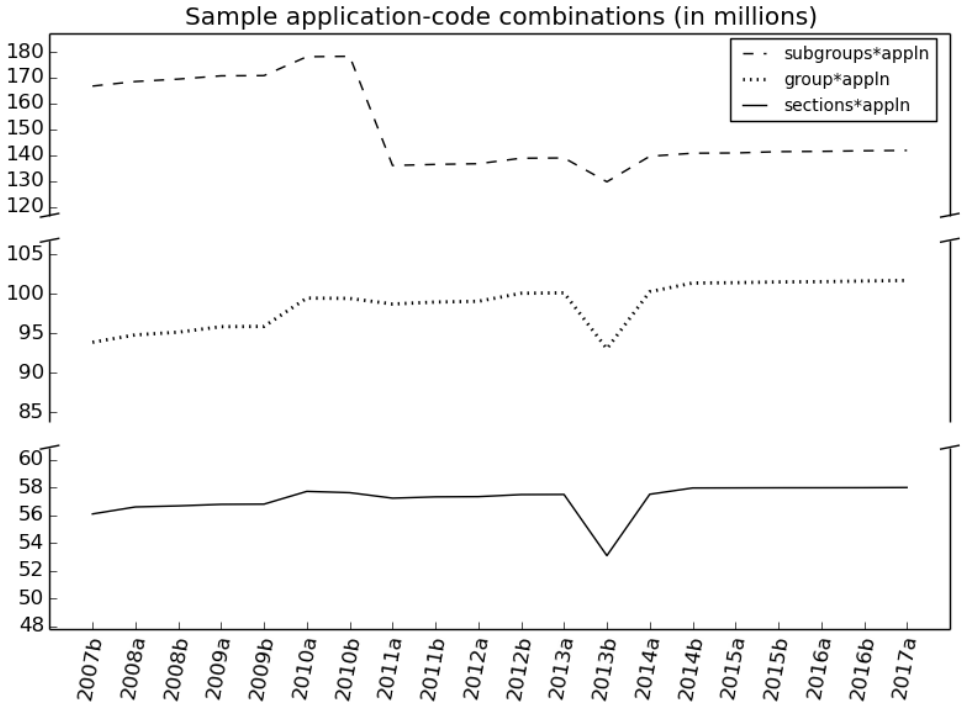


Figure 2: In-sample number of unique (application, IPC code) pairs across PATSTAT versions. Top: IPC subgroups. Middle: IPC groups. Bottom: IPC sections.

event more puzzling is that the drop is not mirrored by a reduction in the number of PATSTAT applications, suggesting that a great number of codes were eliminated after 2010. Up to now, we do not have any information about the causes of such a break in the time series, but the issue is definitely worth exploring. Similarly, there is a (much smaller) drop in the number of both (application, code) pairs and applications between versions 2013b and 2014a. In this case, the plot does not clarify whether the fall in the dashed line is due to primarily to the reduction in the number of applications or to the elimination of IPC codes. Though smaller in magnitude, the second dip is not less puzzling than the previous one. In fact, though it can happen that patent applications and publications are deleted from the master database, it is surprising to see that deletions can outnumber new patent applications. This issue is probably less relevant to reclassification than the previous one, but it still might be important to explore, as, for example, it might help better understand whether it makes sense to investigate reclassification on all patent applications or, rather, one should select the sample with care.

Figure 2 displays the number of unique (application, ipc code) pairs associated to the applications included in all PATSTAT versions. In particular, the top, middle and bottom lines represent the number of unique pairs when IPC codes are defined at the subgroup (all digits), group (8-digit) and section (1 digit) levels respectively. The three plots, which are the in-sample equivalent of the dashed line of figure 1, allow to compare the whole dataset with the subset of patents we concentrate on in this document. Not surprisingly, the increasing trend in the number of (application, code) pairs disappears in figure 1, suggesting that irrespective of its magnitude, reclassification does not affect the volume of data substantially. However, figure 1 also tells us that the main changes in the dataset are reflected in the smaller sample as well, as can be noticed by the fact that the "double dip" of figure 1 is also present in figure 2, especially in the top line. Interestingly, the first large decrease in the number of pairs is mirrored in the top plot of 2 but not in the other ones. This suggests that the large-scale deletion of codes from PATSTAT between versions 2010b and 2011a has mostly affected the number of available subgroup codes, but left the above levels IPC hierarchy almost unaffected in numerical terms. Of course, the graph does not give any information about possible substitutions and the observation is not a hint that reclassification does not take place at higher levels of the IPC. Nevertheless, it is interesting that different dynamics take place a different levels of the classification. Figure 2 also has some interesting implications for the second smaller dip shown in the plots of figure 1. On one hand, since by construction the number of patent applications included in the sample is fixed, it is probably safe to rule out the possibility that the deletion of patents from the database is the leading cause for the transient fall in (application, code pairs) in version 2013b[1]. On the other hand, the small dip, which at a first glance seemed less relevant than the first one, affects the number of unique pairs across the board and thus deserves attention. In fact, we might be able to completely disregard the anomaly affecting version 2011a in future analyses if we choose to focus on higher code aggregations[2]. The same cannot be said for the fall and subsequent rise in the number of pairs around version 2013b, that has to be to reckoned with irrespective of the level of the IPC hierarchy chosen for the analysis.

[1] I would still take this with a grain of salt until we find a way to look into the issue deeper and are able to rule out bugs in the code and other possible perverse effects due to implementation alone.

[2] This is a choice which should probably considered at least in this preliminary phase of the project. In fact, there is a hierarchy within subgroup codes that behaves much like the USPC classification and would probably pose non-trivial technical challenges to extract.

Versions	Dropped codes	Added codes	Applications w. deletions	Applications w. additions
2007b-2008a	461102	2217889	130452	630085
2008a-2008b	482761	1421987	135578	811429
2008b-2009a	413748	1682341	118059	634267
2009a-2009b	393275	480819	114693	166146
2009b-2010a	612517	7923453	416811	2558379
2010a-2010b	341762	503370	145580	176349
2010b-2011a	42949779	681649	27080275	328443
2011a-2011b	27784	455019	20099	240467
2011b-2012a	72881	333433	59612	155012
2012a-2012b	33927	2171120	21448	647932
2012b-2013a	35229	138563	23061	60526
2013a-2013b	9850148	620766	3674508	263200
2013b-2014a	17203	9892337	14108	3618326
2014a-2014b	384810	1569307	236344	992554
2014b-2015a	76415	167242	60340	87064
2015a-2015b	247717	764936	183999	238036
2015b-2016a	79729	184419	63143	86482
2016a-2016b	449867	681211	323074	344176
2016b-2017a	24524	159793	15509	57870

Code changes (codes added or dropped) between versions at IPC-subgroup level.

Versions	Dropped codes	Added codes	Applications w. deletions	Applications w. additions
2007b-2008a	40290	543775	36144	429412
2008a-2008b	41591	119062	37258	113166
2008b-2009a	32879	142175	29792	132076
2009a-2009b	32662	47769	29752	43709
2009b-2010a	1102	921046	1067	843596
2010a-2010b	119525	27007	104103	24503
2010b-2011a	442214	45358	416744	41647
2011a-2011b	3813	97442	3386	93075
2011b-2012a	2475	19054	2283	17220
2012a-2012b	2743	150708	2381	135618
2012b-2013a	3467	12133	2803	10266
2013a-2013b	4423025	17027	3486634	15300
2013b-2014a	4164	4424735	4104	3487207
2014a-2014b	2731	454970	2614	428790
2014b-2015a	2187	9481	2095	8762
2015a-2015b	2446	10016	2362	9034
2015b-2016a	1927	4720	1898	4382
2016a-2016b	6888	11636	6849	10571
2016b-2017a	433	12419	399	10986

Code changes (codes added or dropped) between versions at IPC-section level.

Table 1: Addition and deletion of IPC subgroups and sections.

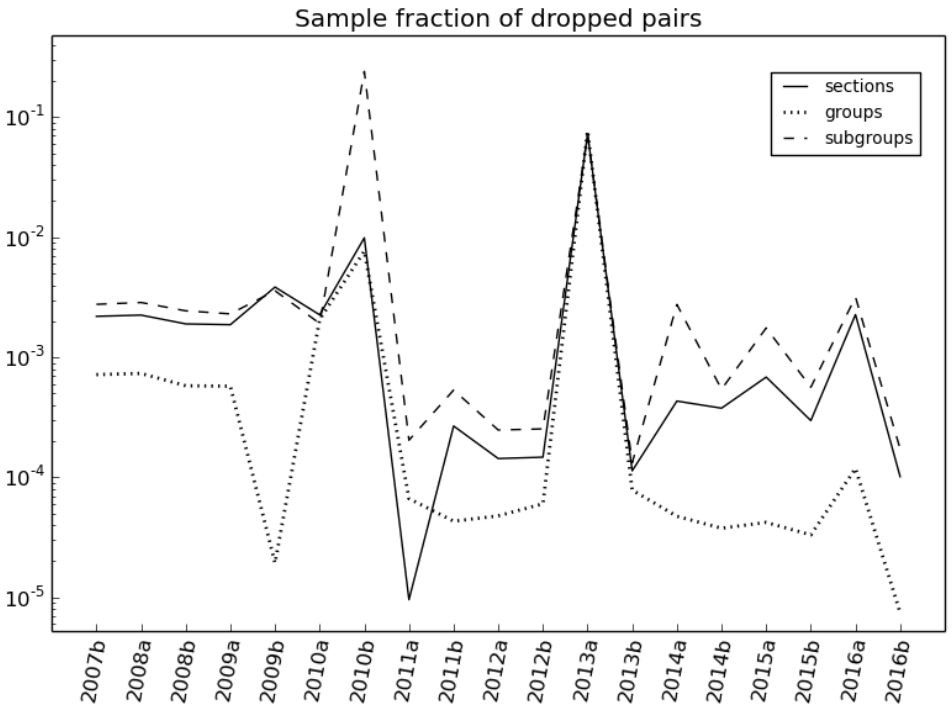
Versions	Dropped codes	Added codes	Applications w. deletions	Applications w. additions
2007b-2008a	206260	1137492	100191	549469
2008a-2008b	213234	555529	103695	421944
2008b-2009a	180791	876224	88614	528865
2009a-2009b	179399	215300	89232	120870
2009b-2010a	369263	3950313	362420	2159718
2010a-2010b	224971	182744	139729	105786
2010b-2011a	984046	276229	812209	171524
2011a-2011b	9478	261233	6791	168995
2011b-2012a	26527	117594	22456	68963
2012a-2012b	14228	1029533	10294	437846
2012b-2013a	14756	70551	8303	37718
2013a-2013b	7286796	214212	3598313	144359
2013b-2014a	10571	7226334	9951	3514610
2014a-2014b	43363	1142862	39584	906617
2014b-2015a	38263	82160	35103	55063
2015a-2015b	69479	161742	65337	107068
2015b-2016a	30204	62854	28448	41540
2016a-2016b	230740	317040	228120	231658
2016b-2017a	10320	82055	9746	41076

Code changes (codes added or dropped between versions) between versions at IPC-group level.

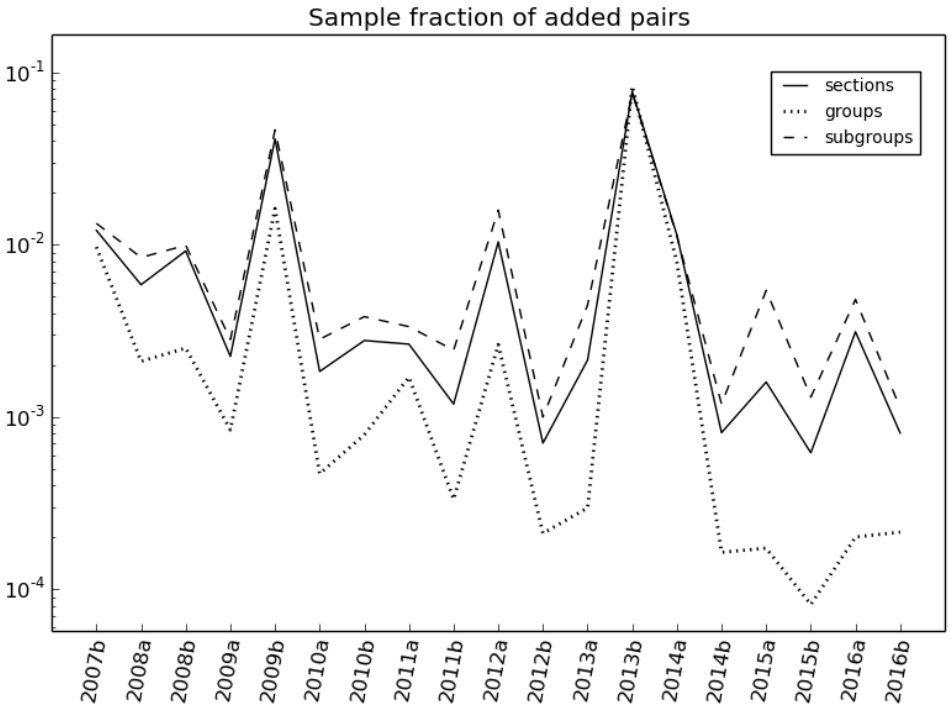
Table 2: Addition and deletion of IPC subgroups and sections.

Table 1 accounts for reclassification defined as the number of (application, ipc-code) pairs that are added or dropped from one PATSTAT edition to the next and reports the number of patent applications that are affected by said reclassification. The table is divided in two parts referring to the two extremes of the IPC hierarchy: subgroups – the most disaggregated level, which includes over 70 thousand codes – and sections – which divide the classification into eight coarse groups. Of course, it is reasonable to expect subgroup codes to change relatively frequently and sections to be more stable. Table 1 details the difference between these two cases and gives an idea of the range of variability within the data. For example, the first row in the top part of table 1 tells that 461102 (application, subgroup) pairs from 130452 applications are dropped between versions 2007b and 2008a, while 2217889 such pairs are dropped from 630085 patent documents. Similarly, the first row of the bottom part of table 1 between versions 2007b and 2008a, 40290 (application, section) pairs are dropped and 543775 are added, affecting respectively 36144 and 429412 patent applications.

Table 2 contains the same information as table 1, but focuses on reclassification at the IPC group level, which is the most disaggregated level of the IPC for which the nested structure is easily extrapolated and thus is probably a good candidate as the level on which to concentrate the analysis. Scrolling down the rows of both tables is informative of the frequency of code additions and deletions over time. Not surprisingly, figures vary quite substantially depending on the resolution at which reclassification is observed. In particular, additions or deletion of subgroup codes are around one order of magnitude more frequent than changes to section-level codes and roughly twice as frequent as group-level codes. It is also interesting to note that tables 1 and 2 display a large fluctuation in the



(a) Dropped (application, ipc code) pairs



(b) Dropped (application, ipc code) pairs

Figure 3: Fraction of (application, ipc code) pairs added and dropped from each PATSTAT version to the next PATSTAT versions.

PATSTAT version	n=1	n=2	n=3	n=4	n=5	n=6	n=7
2007b	46580640	30479	–	–	–	–	–
2008a	47861921	66677	4	–	–	–	–
2008b	48831380	86771	13	–	–	–	–
2009a	49502655	482327	1263	1	–	–	–
2009b	49240566	1663170	5766	17	–	–	–
2010a	49181622	3308872	38954	221	–	–	–
2010b	48630084	4814030	84874	875	2	–	–
2011a	53278299	975707	19561	396	1	–	–
2011b	54460496	1052125	21550	445	1	–	–
2012a	55711992	1393022	45292	1614	22	–	–
2013a	58097702	1504498	49647	1768	24	–	–
2013b	55974359	1760048	80733	4273	169	1	–
2014a	60768487	2051919	100420	4978	194	2	–
2014b	62681091	2322054	125578	7563	497	210	68
2015a	64157821	2533594	139892	8337	536	212	130
2015b	65485138	2911617	189763	12185	764	217	132
2016a	67089428	3222699	212410	13812	916	231	136
2016b	68356841	3710969	246644	18556	1171	244	138
2017a	69675921	3940089	274029	20930	1275	246	141

Table 3: Number of patent applications classified with codes from n versions of the IPC.

number of added and dropped pairs around versions 2011a and 2014a, *i.e.* when the two dips of figures 1 and 2 take place. Moreover, the top part of table 1 responds the most to the first dip – in line with that fact that ipc subgroups are the most affected – while the bottom part of table 1 table 2 are more affected in 2014. Finally, both tables show that the absolute number of added and dropped codes seems to be higher until around 2011 and later decrease. This transition, which seems more abrupt for deletions than for additions, appears also in figure 3, which plots the share of pairs that are dropped and added from a version to the next. The main reason for mentioning the latter feature is that it takes place around version 2011a, which is the first one for which the application identifiers within PATSTAT are guaranteed to be stable across versions. Before that, application identifiers change from version to version, so that they are not directly comparable. This required us to first match patent applications using information about the patents office where they were filed and the number and kind code said office attached to them [1]. Thus, the change in behavior we observe could very well be an interesting feature of the data, but some further testing is probably needed to fully rule out the possibility that what we observe is an artifact due to unsolved issues in the matching procedure.

Tables 3 and 5, which no longer refer only to the data sample we have focused on in the previous exercises but to all of the data, give present a different viewpoint on reclassification by showing that multiple versions of the IPC coexist in each version of PATSTAT and are often used to classify the claims associated to the same patent application. In particular, table 3 shows that the great majority of patent applications use only one version of the IPC, but that a non-trivial share of the documents takes codes from two or more. Table 5, on the other hand, shows that there have been several revisions of the IPC over time, but that the 2006 version is the most employed to classify the bulk of applications. Also notice that table 5 suggests that the fall in the number of (application, ipc,code) pairs we

observe around 2011 and 2014 are mostly due to the deletion of codes belonging to older versions of the IPC. This raises the question whether in general reclassification consists only in substituting deleting old codes and substituting them with newer ones. If it were the case, this might simplify the identification of the set of patents on which to concentrate our efforts and probably reduce the amount of time and storage needed for at least a part of the analyses. A way to test the hypothesis might be to build the analogous of table 5 not for the whole database, but only for the set of patents that appear in all versions and check whether anytime a new version of the IPC appears, the number of codes belonging to older versions stops increasing.

Unsolved issues

First: while the compressed space of all the raw data is just 80GB, the space required to hold all the data in an easy to access form is very demanding. We are waiting to buy a further 8TB hard drive to store the data.

Second: there a minor unsolved issue with application identifiers. As mentioned in the previous section, we matched patent applications from editions prior to 2011a through the triplet (appln_auth, appln_nr, appln_kind) of table TLS201_APPLN, which, according to the official documentation, should ensure that documents are uniquely identified across editions. We found this not to be the case for four triplets, each of which appear identify two slightly different documents. Given that issue involves four database rows out of several million, we simply deleted the rows. It's highly probable that the issue is not concerning, but it might be worth asking next time we speak with the EPO.

Third: we still have not built patent families. Since it turns out that also family identifiers are not guaranteed to be stable across all editions, we think it might be best to wait until we have the full 2017a version and work back from there to reconstruct family structure edition by edition.

Open questions

First: As we mentioned above, we consider the task of linking database versions through a reliable application identifier across versions to be mostly accomplished. However, there is still two issues that might deserve attention. *First*, we found that patent applications are erased from the database between one version and the next and, in a minority of cases (around 20 thousand in total), reappear in later versions. We have not yet explored the issue in detail, but perhaps we should better understand how to treat these *phantom* applications. The *second* finding, which is potentially more concerning, is that the PATSTAT application identifiers, which should be stable across editions from version 2011a onwards, are not actually always stable. We have found that, in a strict minority of cases, the same triplet uniquely identifying a patent application matches different identifiers also when they are guaranteed to be stable (see table 4 for details). Before we proceed too far with the analysis of reclassification, we must be sure that the matching of applications across versions can be trusted, otherwise we might count wrong matches as reclassifications and bias the results of future exercises. We should probably speak with the PATSTAT office about this issue, because it is unlikely to be documented anywhere.

Second: understand whether reclassification only involves deleting or substituting older codes with newer ones. This might have implications for the time and storage requirements to perform future exercises.

Versions	Count
2010b-2011a	55683998
2011a-2011b	3507
2011b-2012a	3147
2012a-2012b	2312
2012b-2013a	3472
2013a-2013b	854
2013b-2014a	6208
2014a-2014b	18817
2014b-2015a	1380
2015a-2015b	3725
2015b-2016a	523
2016a-2016b	1277
2016b-2017a	282

The identifiers should be stable only starting with version 2011a. The first row shows that indeed mismatches are much more frequent before that version.

Table 4: Non matching PATSTAT application identifiers .

Third: according to [2], around 4% of IPC codes in PATSTAT are not related to inventive aspects of the patent application. The figure should be relatively stable across versions. Luckily, non inventive codes are easily identified through the value of field IPC_VALUE included in table TLS209_APPLN_IPC, so in principle they are easy to identify. We did not dig deeper into their analysis because the IPC_VALUE column is not currently indexed (mainly due to storage constraints) and the analysis will require some time. We decided to not delay the report further and eventually add information about non-inventive reports to a future update of this document. We have not excluded non-inventive codes from the exercises of the present document and probably their presence does not make a big difference. However, it might be worth thinking about how to deal with them in the future.

References

[1] EPO. EPO worldwide patent statistical database data catalog - 2014 spring edition. Technical report, 2014.

[2] EUROSTAT. Patent statistics: PATSTAT data quality report. Technical report, 2013.

[3] Joshua Lerner. The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, pages 319–333, 1994.

Version	2006-01-01	2007-01-01	2007-10-01	2008-01-01	2008-04-01	2009-01-01	2010-01-01	2011-01-01	2012-01-01	2013-01-01	2014-01-01	2015-01-01	2016-01-01	2017-01-01
2007a	253825990	58500	–	–	–	–	–	–	–	–	–	–	–	–
2008a	259951898	60342	31980	57400	–	–	–	–	–	–	–	–	–	–
2008b	265506829	60784	35066	74226	13001	–	–	–	–	–	–	–	–	–
2009a	271722822	63585	38132	78878	13247	954781	–	–	–	–	–	–	–	–
2009b	273262191	65898	41434	81328	13423	3581835	–	–	–	–	–	–	–	–
2010a	288021855	114959	61298	103079	15970	6553038	1145542	–	–	–	–	–	–	–
2010b	289875268	118180	64765	108340	16928	6999733	4333555	–	–	–	–	–	–	–
2011a	161958375	121950	68456	115426	17244	750157	808804	494706	–	–	–	–	–	–
2011b	165478976	125320	71366	118042	17583	817200	847204	577331	–	–	–	–	–	–
2012a	169069275	129288	76193	121038	17933	902078	901709	641879	303945	–	–	–	–	–
2012b	174426435	135562	80313	131443	18499	1010236	973746	725925	413475	–	–	–	–	–
2013a	177805720	139281	83653	133655	18741	1072533	1018450	775704	469317	37759	–	–	–	–
2013b	171433273	134287	88559	128964	17651	1189674	1094019	826605	1087105	153197	–	–	–	–
2014a	184810211	148457	93926	139044	19280	1266716	1193822	922175	1208596	452253	3116	–	–	–
2014b	189265220	152786	97394	143537	19616	1422490	1385863	1233234	1308382	785548	70459	–	–	–
2015a	193265494	158708	101438	146599	19988	1524739	1454792	1334138	1431756	892923	218775	8506	–	–
2015b	197081866	166435	106628	163339	21392	1682021	1512547	1498970	1661047	1158782	915128	84381	–	–
2016a	201523199	171507	111219	167246	21800	1781777	1580238	1663672	1762849	1250337	1075917	259268	18304	–
2016b	205434060	176103	115594	171599	22546	1902408	1657519	1777877	2017344	1433946	1157053	298682	577969	–
2017a	208840670	181622	120232	173324	22890	2001217	1711516	1852591	2102141	1525856	1236545	349362	821723	10268

Table 5: Number of patent applications including codes from the IPC version reported in each column

Dynamics of Cumulative Innovation in Complex Social Systems
DCICSS PROJECT

http://opus.bath.ac.uk/55142/9/DCICSS_project.pdf

- **DCICSS Research Team:**
  - Professor Graham Room
    Professor of European Social Policy
    http://www.bath.ac.uk/sps/staff/graham-room/
  - Professor Alastair Spence
    Professor in Numerical Analysis
    http://www.bath.ac.uk/math-sci/contacts/academics/alastair-spence/
  - Dr Evangelos Evangelou
    Lecturer in Statistics
    http://www.bath.ac.uk/~ee224
  - Dr Paolo Zeppini
    Lecturer in Environmental Economics
    http://www.bath.ac.uk/economics/staff/paolo-zeppini/
  - Dr Emanuele Pugliese
    Visiting Fellow at the University of Bath
    Research Fellow at the Institute for Complex Systems, National Research Council (Italy)
  - Dr Lorenzo Napolitano
    Visiting Post-doctoral Fellow at the University of Bath
    Post-doctoral Fellow at the Institute for Complex Systems, National Research Council (Italy)

- **Acknowledgements**
  We acknowledge gratefully the assistance kindly provided by PATSTAT at the European Patent Office, in supplying the datasets we are using for our patent case study.
  http://www.epo.org/searching-for-patents/business/patstat.html#tab1