MAX PLANCK
digital library

PARMA. A full text search based method for matching non-patent literature citations with scientific reference databases. A pilot study.

Technical Report by the Max Planck Digital Library, Big Data Analytics Group

Johannes Knaus and Margit Palzenberger

Munich, February 27, 2018

doi:10.17617/2.2540157

1 Summary

Patent databases contain large amounts of (almost) unstructured references to non-patent literature (NPL). To identify these references is a general research request, as they are an important indicator for determining and quantifying various relationships between science and industry. In the present pilot study, we introduce a *Pa*tent *r*eference *ma*tching method (PARMA) that is able to process a wide range of patent records by using a combination of full text search technology with filtering and matching routines in an RDBMS. Results show that the approach establishes a solid foundation for future analytic studies on the topic.

2 Introduction

According to the definition of the World Intellectual Property Organization (WIPO), a patent is a document that "describes an invention" where invention "means a solution to a specific problem in the field of technology." (World Intellectual Property Organization, 2008, 17). Thus, patents reveal innovations in technical development.

In order to collect technical knowledge and to make it accessible, patents are stored in dedicated databases. Data about patents are valuable not only for examiners of the respective patent offices but also for scientists and engineers who want to get an analytic description of trends, structures and topics in their field. Linking patents with scientific publications provides valuable insights about interactions between academic and industrial research. References to the so called non-patent literature (NPL) are of particular interest here. Existing studies on this topic revealed typical structures and patterns within the data. Van Raan (2017) gives a concise overview of this field of research. Following the fundamental work of Carpenter et al. (1980), he describes the amount of references between patents and scientific literature as a measure of "science intensity" of a technological scientific field (van Raan, 2017, 16, 22), i.e. radically new and disruptive innovations are based on a high number of scientific research publications. Nearly half of the NPL refer to scientific literature that originates from publicly funded research.

NPL citations have been investigated from various different perspectives: Ribeiro et al. (2014) use NPL citations to investigate global innovation networks and transnational transfer of knowledge. Callaert et al. (2014a) point out that indicators of interactions between scientific and technological R&D activities are highly relevant for recent models of innovation systems. These models make it possible to understand the dynamics behind innovations, growth and competitiveness of national economies. Li et al. (2017) evaluate effects of public funding of research on technological development, Veugelers and Wang (2016) identify novel fields of scientific research that are relevant for industry, while Patelli et al. (2017) focus on national effects on R&D.

Despite the importance of NPL data for thorough analyses of the interrelations between scientific publications and patents, the major part of the NPL references is available only as unstructured free text data. They do not follow any uniform formatting rule and appear in a large number of variants. Repeatedly citations are incomplete or inconclusive. This hinders a matching of references from patent databases with those from scientific publication databases like Web of Science (WoS) or Scopus (SCO). As a result, the majority of studies on NPLs was restricted to relatively small sample sizes. Coward and Franklin (1989) identified 255 patent-paper pairs in 2,452 patents by matching proper names of persons and institutions between inventors and publication authors. Boyack and Klavans (2008) identified NPL authors that also occur as inventors in 56,000 US patents between 2002 and 2006. He matched rare names only. Magerman et al. (2015) use text mining and content similarity techniques to analyze references from 88,248 patent documents from the European Patent Office (EPO) and the US Patent Office (USPTO). Li et al. (2017) match scientific publications funded by the US National Institute of Health (NIH) to investigate effects of science funding on the generation of patents in chemistry and biomedicine. The only large scale studies published so far have been conducted by Shirabe (2014) (15M NPLs from USPTO) and Callaert et al. (2014b) (11M NPLs from WIPO, EPO and USPTO).

The current study pilots a framework for an efficient and reliable matching of complete NPL sets from large patent offices to global scientific literature databases. To cope with the high number of potential comparisons and the missing structure of the NPL entries, we explored a pipeline that combines Solr full text search engine technology with SQL pre- and postprocessing. The setup developed was run on 22M unique NPLs from WIPO, EPO and USPTO as citing sources against 50M Web of Science records as cited targets.

3 Data Sources and Methods

3.1 Data Sources

The main data sources for this study were DOCDB, the core reference raw data product of the European Patent Office, Web of Science (WoS), a scientific publication database by Clarivate Analytics and JUNE, an MPDL in-house database for journal metadata (see table 1). The goal of the study made it necessary to aquire commercial licenses for DOCDB and WoS in XML formats.

DOCDB (European Patent Office, 2017b) is provided as XML data in two ways, as a yearly snapshot called "backfile" containing all patent data up to the snapshot creation date and as biweekly updates called "frontfiles". For this piloting study we focused on the most recent backfile (7554 data files with a total file size of 715GB). The according XSD files are publicly available on the EPO website (European Patent Office, 2017a).

DOCDB includes records from more than 100 patent offices. For the pilot study we decided to focus on records assigned to the three major western patent offices, i.e. World Patent Office (WIPO), the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). Table 2 gives an overview of record numbers in the three subsets. The majority of records in these subsets is registered in English and thus compatible to the metadata records of Web of Science.

Since August 2016, DOCDB includes "rich" citation data (i.e. EP search report citations) have been added which contain comprehensive citation data with entries split into subfields. However, only a subset of the references, mostly from EPO and WIPO, is available in rich format and all information is contained in the standard citation field as well. Therefore, these extra XML fields are not taken into account here.

Web of Science (WoS) XML data are licensed via the German Competence Centre for Bibliometrics (CCB) and provisioned by FIZ Karlsruhe as a relational Oracle database. The dataset encompasses the journal core collection (SCI, SSCI and AHCI) as well as conference proceedings from 1980 onwards. The raw data are routinely processed by MPDL to feed a PostgreSQL data warehouse for analytical purposes. Thereby the data are partially cleaned and enriched with standardized data for journals and selected institutions.

MPDL JUNE serves in-house needs for journal metadata standardization. All Web of Science journals are covered. In the context of this study it was used to enrich journal names as used by Web of Science by additional name variants.

Name	Provider	Format	Coverage		Size	base records
			starting	release date		
DOCDB (backfile)	EPO	XML	1782–	2017-06	715GB	~100M
Web of Science (core collection)	Clarivate	XML	1980–	2017-01	431GB	~50M
JUNE	MPDL	RDB	1980–	2017-05		~80K

Table 1: Data source properties

Object type	Set	All	WIPO	EPO	USPTO
Patent records	total	101.3M	3.7M	5.8M	14.8M
NPL references	total	36.7M	4.6M	4.6M	24.3M
	unique	24.4M	3.5M	3.5M	15.0M

Table 2: DOCDB data volumes

3.2 Methods

For the present study, we decided to use Solr, a search engine technology based on Lucene (Apache Software Foundation, 2017), to provide fast and comprehensive string matching functions. Standard settings, however, would not have been powerful enough to cope with the large amount of data that had to be compared (more than 20M by 50M pairs). Extensive tests of various Solr setups were run to find a good enough setting with the technical equipment available (20 cores, 130GB RAM, SSDs). The final configuration provided us with a throughput of 35 sec/1000 queries, which was considered sufficient for the needs of the pilot study.

The Solr full text search procedure finds a ranked set of potential NPL-WoS matches, however, it does not allow for an absolute measure to decide between true and false positives. These decisions were made by field-wise SQL comparisons of the top search results of each query.

The complete pipeline consists of six main steps (fig. 1):

Step 1 DOCDB XML Extraction:
Extraction of relevant nodes from the XML raw data and import of the extracted information into a PostgreSQL relational database

Step 2 SQL Preprocessing:
Filtering and cleaning of data, extraction of publication years and other patterns Generation of full text query strings from the NPL entries.

Step 3 Solr Search Index:
Export of relevant fields of Web of Science into a Solr search index

Step 4 Solr Queries:
Search of the NPL full text strings (step 2) in the Web of Science index (step3)

Step 5 SQL Postprocessing:
Field-based scoring of the top ranked NPL–WoS pairings

Step 6 Quality Assurance:
Manual processing of random samples to verify categorization and matching

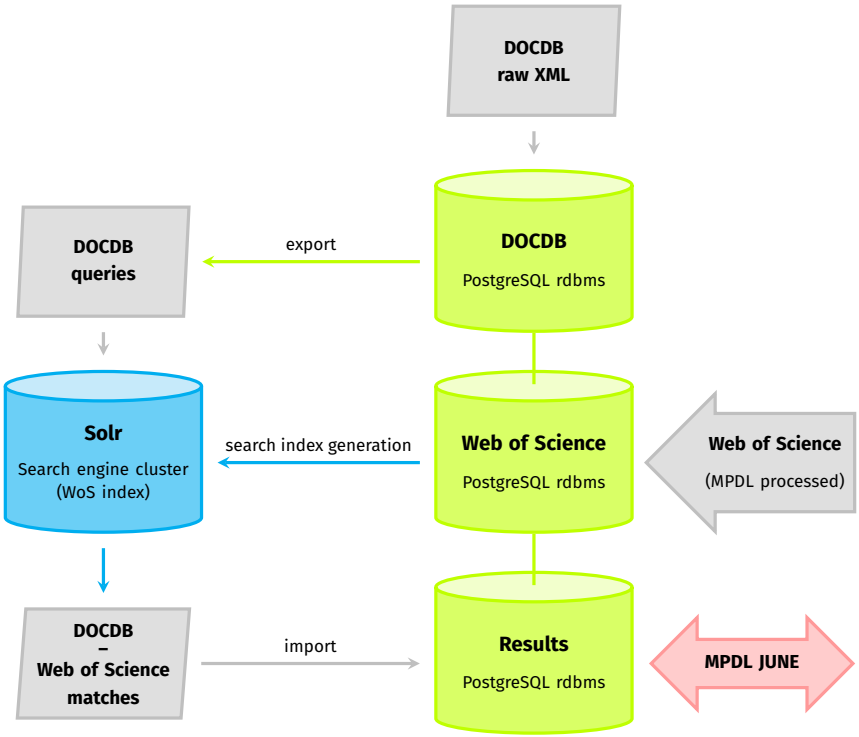


Figure 1: Processing pipeline

3.2.1 DOCDB XML extraction

We extracted information from the 7,554 DOCDB XML files and transferred it to a PostgreSQL database. Based on the official documentation and the XML Schema Definitions, we selected 39 elements (nodes or attributes) relevant for our project. The contents of these elements were extracted by a Python script iterating over the records and accessing the data with appropriate XPath expressions. In total 101,308,828 records were extracted and written into a set of CSV files. The latter were imported into the PostgreSQL database, where they were cleaned, aggregated and prepared for the matching process. The fields "country" for the patent offices and "nplcit" for the reference entries where those of special interest for the analyses presented. The "nplcit" entries were transformed to lower case and the unique variants of all entries across the database were used for further processing.

3.2.2 SQL preprocessing

An initial inspection of the NPL references in the DOCDB subset selected showed an enormous variance. The same targets are registered in multiple variants and thus the number of unique references is only slightly lower than the total number. A substantial amount of the entries does not even refer to what we would consider non-patent-literature. We found records that are cross-references to patent databases or describe steps from patent processing. We also found records that are very short and contain cryptic or incomplete entries that cannot be reliably matched to literature targets. Random examples can be found in the list of quality assurance records provided in the supplementary material.

To cope with this situation and optimize the subsequent Solr query performance, several preprocessing steps were executed. They all were based on PostgreSQL regular expressions combined with full text indices.

As the Solr matching ties up substantial processing time, we tried to lower the number of records to be processed. This was done by filtering out those records that can be identified with sufficient certainty as non-targets. Two main groups were excluded from further general processing:

The category "**most likely no NPL references**" includes entries that do not refer to literature or literature databases. This comprises references to other patents within DOCDB, to other patent databases, and references to patent processing steps. In general, cross-references have rather clear patterns, that can easily be detected. This is however not true for the subset that reflects patent processing steps. These can be found in a substantially high number of variants and thus were covered with limited recall only.

Our target database Web of Science includes only a selected subset of journal articles and conference proceedings. We therefore defined a category that comprises records that are "**most likely non-targets**". References with a publication year before 1980 could be detected with sufficient precision and were included into this category. Entries with a length of less than 20 latin characters were also added to this category.

The remaining "**candidates for matching**" were cleaned from special characters and substrings that certainly will not be found in the Web of Science index, as for instance URLs and XP-numbers.After cleaning, all candidates from this group were used as Solr query strings. To allow for the blocking concept, the numbers that are likely publication years were extracted and used as routing parameter in the Solr queries (see step 3 and 4).

Amongst the candidate strings we tagged the category "**probably non-targets**" (books, reports, literature and fact databases, norms, etc.). These, however, could not be identified with sufficient precision and recall and thus remained included in the matching set.

3.2.3 Solr Search Index

The core of this routine is a search engine cluster which operates on a searchable full text index that was created by collapsing the most relevant fields of the Web of Science database (cf. XML-listings in the supplementary material):

- Web of Science identifier (UT)
- digital object identifier (DOI)
- publication year (PY)
- source title (SO)
- source short title (J2)
- source volume (VL)
- source start and ending page (BP, EP)
- source number (ARTN)
- up to 3 authors (lastname, firstname(s))
- publication title (TI)

In a first step, we split up the search index into several smaller parts, a technique that is called **sharding**. Sharding allows to search an index in less time as all shards can be searched in parallel. Since release 4.0 in 2012 Solr offers SolrCloud, a feature allowing to set up and manage distributed search and indexing that is scalable and easy to maintain.

Sharding alone does not provide us with a sufficient search speed to effectively process the NPL queries, as still the

Parameter	Explanation	Values
defType	query parser	eDisMax
mm	minimum should match parameter	3
qf	indexed field to be searched	WoS full text field
fl	fields to return as result	unique id, score
route	shards to be searched	publication year(s) extracted
q	main query	NPL full text record

Table 3: Solr Query Parameters used

whole index would have to be searched for every query. To avoid this, we combined sharding with **blocking**, i.e. we arranged the shards into blocks formed by a previously known information entity. We chose the publication year as the blocking factor, as it is relatively easy to extract from the NPL records (see step 2). On the SolrCloud side, we defined an "implicit" router, which allows it to use a router field parameter during the creation of a "collection" (i.e. the cluster of indexes). The router field is an additional field in the data which carries the name of the shard. During indexing, each "document" (Web of Science record) is automatically routed to the specific shard defined by the router field. This way, we spread the index over 38 shards with one shard per publication year. Figure 2 schematically shows a part of the index. Shards are gathered into one collection and are replicated over several Solr nodes, which in the present set up are Jetty/JVM instances running on the Server under separate individual ports.

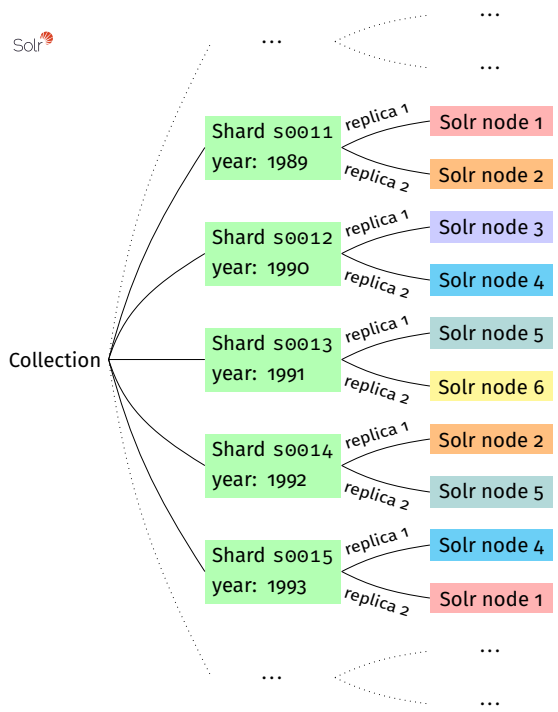


Figure 2: SolrCloud with a sharded index blocked by publication year

3.2.4 Solr Queries

The actual query process is realized by a Python script that sends search queries to the REST (representational state transfer) interface of the SolrCloud cluster, collects the Solr responses, and saves the resulting data into CSV files. To increase performance, query processes run in parallel using multiple CPU cores.

The Solr query parameters given in Table 3 yielded optimal results with regard to the requirements of the project. Apart from obvious choices like the query string itself, index field queried, and the format of the result, we made some specific decisions. The extended Maximum Disjunction (eDisMax) query parser allows google-like full text searches and the use of further parameters that restrict the set of results. In our case, this was primarily the Minimum Should Match Parameter (mm) which we set to 3, i.e. a minimum of three clauses/terms should match in order to retrieve a result.

The route parameter is relevant for the Solr index set up chosen (step 3). With a blocked and sharded index, the search requests sent to the SolrCloud server are automatically routed to the shard that contains the documents with the referring publication year, such that only this shard needs to be searched.

Solr was set to retrieve the top 10 records ranked by the eDisMax scoring algorithm. This score takes into account term frequency, inverse document frequency and inverse field length. It is a relative measure, that applies to a given query (= NPL reference) only. Identifier, scores and runtimes of the top 10 matches were written to CSV files and imported into the PostgreSQL databases.

3.2.5 SQL Postprocessing

Whenever the criteria set for the Solr matching function (tab. 3) are met, Solr retrieves records from the index. Due to our very liberal settings the top records retrieved include a high fraction of false positives. Solr ranking scores cannot be used for further selection as they are not comparable accross queries and they cannot be interpreted as an absolute measure of matching quality. Therefore

we used SQL procedures to calculate a domain specific and absolute matching score for the NPL–WoS pairings suggested by Solr. For the pilot study, we restricted these analyses to the top ranked pairing for every NPL record.

The matching fingerprint is made up from individual comparisons for six fields of the WoS records (fig.3): publication year (PY), source volume (VL), source beginning page (BP), first author (AU), article title (TI), and source title variants (SO). For all fields, the exact strings were searched within the preprocessed NPL string. When found, they were marked with an 1 in the scoring fingerprint. Article and source titles where specifically processed as there is a high probability that they do not match exactly between the NPL reference and the WoS record.

Article titles: many NPL records delimit the article title with double quotation marks. These parts were extracted and the relative Levensthein distance to the paired WoS article title was calculated. If that distance was below a threshold of 10% of the average string lengths the scoring position was marked 1.

Source titles: source titles show an enormous variance within the NPL records. Any possible (and impossible) abbreviation is used and there is no pattern frequent enough to use it for a reliable extraction of the source title from the full string. MPDL maintains an in-house journal databases (MPDL JUNE) including all WoS journal and series titles and their standard abbreviations. These, however, did not cover even frequently used abbreviations in NPL records. Therefore, we extracted potential source titles with very broad regular expressions (e.g. anything between a dot and 'vol'). The most frequent strings derived were then matched against the existing journal titles by a semi-manual procedure. All variants derived from this procedure were used for the scoring comparison.

The total matching score was calculated by the sum of hits in the fingerprint and thus ranged from 0 (no WoS record with at least 3 common terms found) to 6 (all WoS elements checked have been found in the NPL string).

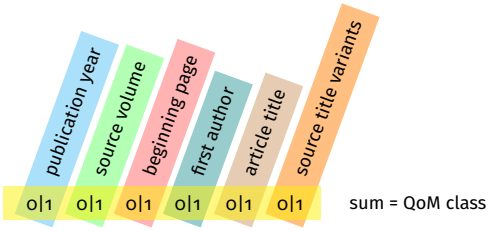


Figure 3: Matching Fingerprint

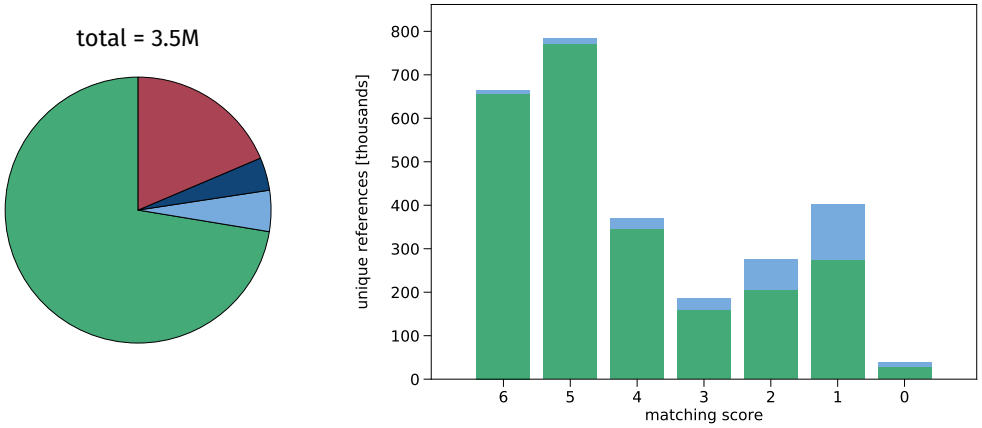
3.2.6 Quality Assurance

For quality assurance, we randomly selected 1000 references per patent office.

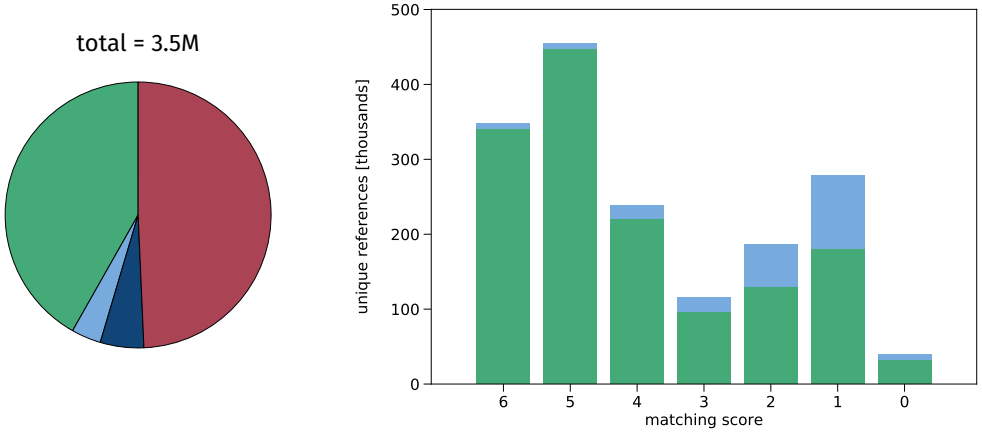
Every record was inspected by an expert and qualified as "target" (part of our Web of Science data set) or "no target" (not part of it). Non-targets where either identified directly in obvious cases or in case of doubt verified by searching the Web of Science web interface with a strategy optimizing recall.

For those records that had been identified as target the Solr mapping between DOCDB and Web of Science was categorized into "wrong" and "found".

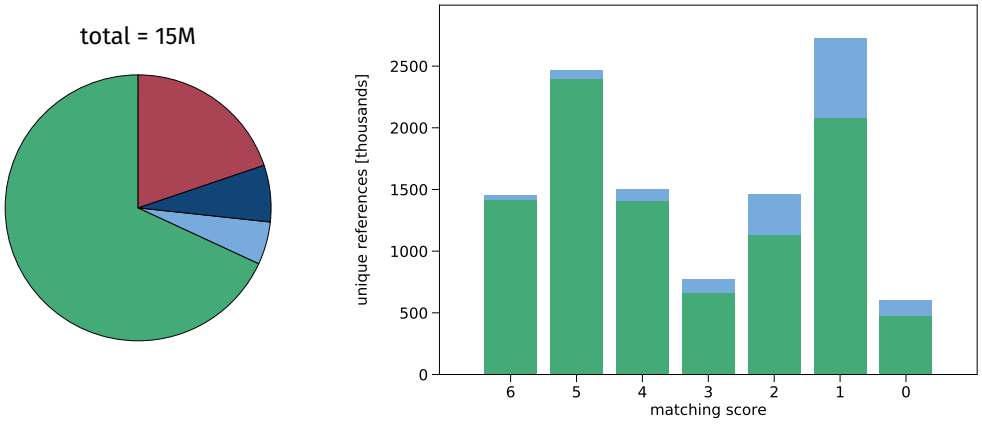
World Intellectual Property Office (WIPO)



European Patent Office (EPO)



United States Patent and Trademark Office (USPTO)



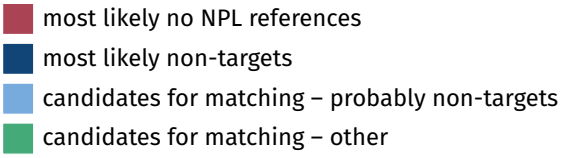


Figure 4: Categorization of NPL references and scoring results.
score = number of matching fields; 0 = no match, 6 = best match

		WIPO				EPO				USPTO			
			manual classification [%]				manual classification [%]				manual classification [%]		
		n	no targets	targets		n	no targets	targets		n	no targets	targets	
processing category				wrong	found			wrong	found			wrong	found
candidates	score 6	195	0	0	100	112	0	0	100	99	0	0	100
	score 5	215	0	0	100	133	0	0	100	161	0	0	100
	score 4	103	4	2	94	57	0	2	98	88	5	1	94
	score 3	45	36	13	51	35	34	6	60	44	36	9	55
	score 2	73	89	10	1	47	83	13	4	99	85	8	7
	score 1	104	98	2	0	74	95	5	0	204	96	4	0
	score 0	12	100	0	0	5	100	0	0	44	100	0	0
no target	short	0				2	100			5	100		
	pre 1980	41	100			54	100			62	100		
no NPL	derwent	29	100			18	100			4	100		
	see references	170	100			429	100			0			
	patent processing	13	92			34	100			190	99		
		1.000				1.000				1.000			

scoring precision		n	cum. [%]	n	cum. [%]	n	cum. [%]
	>99%	422	56	250	54	304	41
	>90%	207	84	131	82	292	81
	>80%	73	94	47	92	99	94
	<80%	45	100	35	100	44	100
		747		463		739	

Table 4: Manual quality assurance

4 Results

Our pilot setup made it feasible to process all matching candidates assigned to the three patent offices chosen. In total 21.7M records were handled by the methods described with a total machine processing time of less than 8 days. The lion's share of that time was allocated to the Solr searching process, which has to select potential matching candidates out of 50M Web of Science records for every NPL record. Several measures increased performance at that step: Preprocessing filtered out most likely non-candidates and shortened reference strings and thus the number of matching terms. The Solr setup allowed to efficiently run several tasks in parallel. The far most effective measure turned out to be the implementation of blocking, in our case the creation of separate Solr indices per publication year.

Regular pattern analysis revealed substantial insight into the types of entries and the quality of our matching procedure. During preprocessing, the patterns successfully identified a large volume of non-target entries with sufficient certainty to exclude them from the matching process. We also were able to extract potential publication years in most references, a prerequisite for our blocking approach. As foundation for the scoring process, potential document and source titles were extracted by regular patterns.

Automated bulk processing was backed by manually processing random samples of 1000 records for each patent office. This gives valuable insight into recall and precision of our procedures. The quality assurance set comprises examples for all major variants and thus is included in the supplementary material.

Figure 4 and table 4 give an overview over the volumes of basic categories (pie charts) and scoring classes (bar charts) along with the estimated accuracy (manual quality assurance) for the three patent offices. For the basic categories defined in 3.2.2, we find the following volumes:

"**most likely no NPL references**" (red segments in figure 4): Around 20% of WIPO and USPTO and 50% of the EPO unique records were identified as entries different from what we would consider non-patent-literature references. In WIPO and EPO subsets, these include a very high fraction of simple cross-references to other patents or patent databases. These cross-references were determined by simple regular expressions with sufficient accuracy. This is more challenging for USPTO, as the reference fields of this office include many strings that seem to be annotations out of the patent handling process. These strings are highly variable and patterns show a typical long tail distribution. Even with considerable effort in the development of adequate regular expressions these could be captured with limited recall and precision only.

"**most likely non-targets**" (dark blue): References to publication years between 1900 and 1979 ("pre 1980") make

up between 4 and 7 percent of the unique references in the three offices. They were identified with high reliability (> 99%, table 4) and thus excluded from further processing.

"**probably non-targets**" (light blue): There are many journals, conference proceedings, books, reports, literature databases, and other source types, that are not covered in the Web of Science core collection. We put some effort to find appropriate patterns for these, but within the scope of the pilot study this was limited to the more prominent cases. We identified some 5% for each office via the regular expressions developed. This recall is far from complete. Manual quality assurance suggests this segment to be approximately 15% for EPO and 20% for WIPO. The even higher share of around 35% for USPTO includes also undetected non-NPL-entries. The complete category remained part of the candidates for matching.

The "**candidates for matching**" (green and light blue) account for 2.7M (77%) in WIPO, 1.6M (45%) in EPO and 11M (73%) in USPTO. The matching scores for the three offices showed similar frequency distributions. Medium scores are less frequent than those at both ends of the scale (with the exception of score 0, which labels the rare cases where no Solr match was retrieved).

Manual quality assurance shows that scores of 5 and 6 reliably indicate the correct match of a NPL reference to a Web of Science article. The error rate for these scores is less than 1%. A score of 4 still shows an acceptable precision with an error rate of less than 10%. Scores of 0 and 1 have a high prevalence of non-target references and some rare targets that were not matched to their correct counterparts. Thus, these matches can safely be excluded from any further consideration.

The transitional zone between these two rather well-defined clusters is found for the scores 2 and 3. Matches with these scores have the least reliable predictive power. These records need to be excluded from any analysis that cannot account for false positives. They still include many targets and thus any refinement of the method has to lower their share.

Summing up manual quality assurance, we achieved the highest reliability (precision > 99%, scores 0 for non-target, 5 and 6 for targets) for more than half of the matching candidates in WIPO (56%) and EPO (54%) and little less for USPTO (40%). If we accept error rates of up to 10%, i.e. if we include score 1 for non-targets and 4 for targets, we can cover more than 80% of the records in any of the three offices.

In short, successful matching was achieved for more the 6 million unique NPL-references with the highest reliability (scores 5 and 6) and another 2 million with an expected error rate of less than 10 percent (score 4).

5 Discussion

The framework implemented for the pilot study fulfilled our goal to process a significant amount of NPL references with acceptable matching quality. Current hardware and mature full text search and database technology enabled us to run the pipeline with acceptable throughput. Building on this, we can envisage further development including new data sources and optimizing algorithms.

Previous studies report issues with respect to processing time. Shirabe's (2014) algorithms run "a couple of weeks" for 15M records (USPTO 1992-2012). With the current setting, we are able to process this volume in about a week. The lion's share of that time is consumed by Solr full text matching. The configuration for this step is ready to be scaled for faster hardware, extended memory and increased parallelization without major changes. The key feature of our concept is the separation of the Solr full text matching from single field comparisons and string pattern analyses implemented in the PostgreSQL database. The development of the latter needs many more cycles and thus benefits considerably from practicable stepwise processing times in the range of minutes to a maximum of some hours for the full data set.

Callaert et al. (2014b) matched a similar set of references from the EPO PATSTAT (a derivative of DOCDB) to Web of Science. They processed 12.5M NPL records of the patent application years 1993 to 2009 from the offices WIPO, EPO, and USPTO. Using a specifically trained machine learning algorithm (Callaert et al., 2012), they identified 52% as "scientific" NPL references, including citations to journals and conference proceedings. These references were further processed in two steps using single field comparisons with high recall analogous to our Solr full text matching and a third step with field based filters to improve precision similar to our SQL scoring. With that procedure 3.3M valid matches were found, i.e. 26% of the total number of NPL records. This compares well to the 28% yield for our best scores 5 and 6.

With the current setting we found 8 million matches with satisfactory precision (scores 4-6) out of the 22 million unique references analyzed. If we extrapolate results from our manually processed random samples, we can expect another 1.5 million to match with Web of Science records. We suppose that most of these have been successfully matched by Solr but lack confirmation by the scoring process.

There are many tweaking options for an improvement of the scoring algorithms: For the pilot, we analyzed the top ranked Solr result only, but there might be cases where a lower rank gives a better score. A substantial gain in precision could be achieved by an identification of further source title variants. Eventually, the inclusion of further author names also would be an additional benefit. For the reliable exclusion of non-targets, there are many options to improve the detection patterns. However, all measures that include work on regular expressions, are typical long-tail problems and any increase in discriminatory power is accomplished by an exponential increase in development time. Future studies might exploit the input of machine learning algorithms at these steps.

Another approach to increase the fraction of reliable matches would be an extension of the target database. We already run a successful pilot on Scopus, another global literature database, that has a broader coverage than Web of Science. Crossref, the global DOI registry, would also be a very valuable target resource. Beyond that, more specialized collections like ArXive, BIOSIS, CiteSeer and many others could be included. The pilot study suggests that our framework is ready to cope with these scenarios.

Acknowledgments

This study was made possible by data made available by the German Competence Centre for Bibliometrics (cf. website) which is supported by the German Federal Ministry of Education and Research (BMBF) under grant number 01PQ17001. We gratefully acknowledge the helpful conversations with D. Harhoff, F. Gaessler, and D. Risteski within our joint project with the Max Planck Institute for Innovation and Competition, the advise of G. Weikum of the Max Planck Institute for Informatics, the indispensable work in manual quality assurance of our student assistents N. Nguyen and L. Scheibenreiff, and the Max Planck Computing and Data Facility (MPCDF) for providing the necessary server infrastructure.

References

Apache Software Foundation (2017). Solr. Open source enterprise search platform built on Apache Lucene™. Version 6.4. URL https://lucene.apache.org/solr.

Boyack, K. W. and Klavans, R. (2008). Measuring science–technology interaction using rare inventor–author names. *Journal of Informetrics*, 2(3):173–182. doi:10.1016/j.joi.2008.03.001.

Callaert, J., Grouwels, J., and Looy, B. V. (2012). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, 91(2):383–398. doi:10.1007/s11192-011-0573-9.

Callaert, J., Pellens, M., and Van Looy, B. (2014a). Sources of inspiration? Making sense of scientific references in patents. *Scientometrics*, 98(3):1617–1629. doi:10.1007/s11192-013-1073-x.

Callaert, J., Vervenne, J.-B., Van Looy, B., Magerman, T., Song, X., and Jeuris, W. (2014b). *Patterns of Science-Technology Linkage*. Publication Office of the European Union, Luxembourg. doi:10.2777/55249. OCLC: 931584316.

Carpenter, M. P., Cooper, M., and Narin, F. (1980). Linkage Between Basic Research Literature and Patents. *Research Management*, 23(2):30–35. doi:10.1080/00345334.1980.11756595.

Coward, H. R. and Franklin, J. J. (1989). Identifying the Science-Technology Interface: Matching Patent Data to a Bibliometric Model. *Science, Technology, & Human Values*, 14(1):50–77. URL http://www.jstor.org/stable/689670.

European Patent Office (2017a). DTD/Schema repository. Version 14.7 – DOCDB. URL https://publication.epo.org/raw-data/product?productId=91.

European Patent Office (2017b). Exchange Format. EPO - Patent Information Resource. Exchange of Patent Information as produced by the EPO from their master documentation database DOCDB. Version 2.5.7. URL http://documents.epo.org/projects/babylon/eponet.nsf/0/6266D96FAA2D3E6BC1257F1B00398241/$FILE/T09.01_ST36_User_Documentation_vs_2.5.7.2_en.pdf.

German Competence Centre for Bibliometrics (2018). Official website. http://www.bibliometrie.info. Accessed: 2018-02-10.

Li, D., Azoulay, P., and Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, 356(6333):78–81. doi:10.1126/science.aal0010.

Magerman, T., Looy, B. V., and Debackere, K. (2015). Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology. *Research Policy*, 44(9):1702–1713. doi:10.1016/j.respol.2015.06.005.

Patelli, A., Cimini, G., Pugliese, E., and Gabrielli, A. (2017). The scientific influence of nations on global scientific and technological development. *Journal of Informetrics*, 11(4):1229–1237. doi:10.1016/j.joi.2017.10.005.

Ribeiro, L. C., Kruss, G., Britto, G., Bernardes, A. T., and da Motta e Albuquerque, E. (2014). A methodology for unveiling global innovation networks: Patent citations as clues to cross border knowledge flows. *Scientometrics*, 101(1):61–83. doi:10.1007/s11192-014-1351-2.

Shirabe, M. (2014). Identifying SCI covered publications within non-patent references in U.S. utility patents. *Scientometrics*, 101(2):999–1014. doi:10.1007/s11192-014-1293-8.

van Raan, A. F. (2017). Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. *Journal of Data and Information Science*, 2(1):13–50. doi:10.1515/jdis-2017-0002.

Veugelers, R. and Wang, J. (2016). Novel science for industry? In: *2016 IEEE International Conference on Management of Innovation and Technology (ICMIT)*, 270–274. doi:10.1109/ICMIT.2016.7605046.

World Intellectual Property Organization (2008). *WIPO Intellectual Property Handbook Policy, Law and Use*. WIPO, Geneva, 2nd edn. URL http://www.wipo.int/edocs/pubdocs/en/intproperty/489/wipo_pub_489.pdf.

Supplementary material is available under
http://pure.mpg.de/pubman/item/escidoc:2540157

- Excel document containing raw data and example regular expressions
  mpdl_rio_parma_techrep_201802_supp01.xlst
- CSV file with quality assurance data (section 3.2.6)
  mpdl_rio_parma_techrep_201802_supp02.csv
- XML files containing the relevant Solr indexing and configuration snippets
  mpdl_rio_parma_techrep_201802_supp03.xml
  mpdl_rio_parma_techrep_201802_supp04.xml