

Bericht für das Kompetenzzentrum Bibliometrie

HIERARCHISCHE ARTIKELKLASSIFIZIERUNG

Von Haiko Lietz und Arnim Bleier

Köln, 21. Oktober 2015

ZUSAMMENFASSUNG

Das Klassifikationssystem, das im *Web of Science* verfügbar ist, basiert auf der Einteilung von Fachzeitschriften in relativ starre und wenige Klassen, die stark an historische Vorstellungen von Disziplinen und Grenzen von Domänen angelehnt sind. Es ist zur Untersuchung von Forschungsfronten und -gebieten, die quer zu solchen Einteilungen liegen, nicht geeignet.

Aufbauend auf Vorarbeiten der Partner im Kompetenzzentrum wird durch ein probabilistisches Verfahren des unüberwachten maschinellen Lernens eine zitationsbasierte Artikelklassifikation vorgestellt. Hierfür ist ein Standardverfahren (Latent Dirichlet Allocation) skalierbar gemacht worden, um es auf große Datensätze anwenden zu können.

Zunächst wird der vollständige Jahrgang 2010 in 6 Wissenschaften, 39 Disziplinen und 250 Felder eingeteilt. Die identifizierten Cluster sind bedeutungsvoll und gut interpretierbar. Sie werden durch Keywords beschrieben und evaluiert, indem sie der etablierten Zeitschriftenklassifikation gegenübergestellt werden. Wie durch die Methode erwartet, befinden sich die gefundenen Themengebiete im Überschneidungsbereich klassischer Abgrenzungen, d.h., sie bilden oft multidisziplinäre Forschungsgebiete ab.

Die probabilistische Natur der verwendeten Methode ist besonders geeignet, Publikationen für bibliometrische Analysen zu definieren, da Publikationen immer mit bestimmten Wahrscheinlichkeiten einem Thema angehören und Themen zu bestimmten Wahrscheinlichkeiten aus zitierten Referenzen bestehen. Sind Themen bzw. Publikationemengen für bibliometrische Studien noch nicht bekannt bzw. definiert, lassen sie sich mit verhältnismäßig geringem Expertenaufwand am Anfang von Studien festlegen.

Da Klassifikationen wie die hier präsentiert erheblich von der gewohnten Zeitschriftenklassifikation abweichen, sind Akzeptanzschwierigkeiten denkbar. Im Gegensatz zu expertenbasierten aber statischen Einteilungen der Wissenschaft ermöglichen sie jedoch organischere Abbildungen von tatsächlichen Forschungsprozessen. Um eine arbeitsfähige Klassifikation zu erstellen und unseren Auftrag in Gänze zu erfüllen, werden wir nun den Gesamtbestand ab 1990 klassifizieren. Ziel ist es dabei, eine fünfstellige Anzahl Themen zu identifizieren.

Worauf es zu achten gilt, ist hiermit identifiziert. Das Verfahren braucht ausreichend Zeit um zu konvergieren, und der Speicherbedarf muss reduziert werden, wahrscheinlich durch eine hybride Klassifizierung, die auf Zitationen und Wortverwendungen basiert.

1 INTRODUCTION

The databases of the Competence Centre for Bibliometrics (KB) offer journal-based classification systems (the so-called subject categories). This classification system has its origin in expert judgement. This supervision comes at the cost of not being automated. The problem with such supervised/unautomated classification systems has been identified. (a) Experts can only define classes that they have observed, but observation is preconditioned by what has been observed in the past. Systems are historical and not suited to identify emerging topics. (b) Experts can only define a limited number of classes while keeping them somewhat distinct. Systems are not suited to identify fine-grained specialties at subfield levels.

Two KB projects have aimed at classifying publications. Sahrhage et al. (2013) use all available meta data (authors, cited references, subject categories, conferences, journals, and keywords) in a graph-theoretical framework but, due to scalability issues, are not able to move beyond clustering the discipline of mathematics. Mund (2014) uses words, cited references, and authors in an unsupervised topic modeling setting and is able to identify emerging topics in a small sample of 3,271 articles from 2007.

Here, to create a useful classification of publications in the KB databases, we employ Latent Dirichlet Allocation, the method also used by Mund (2014). While graph-theoretical approaches are frequentistic and typically result in publications either belonging or not belonging to a class, Latent Dirichlet Allocation is a Bayesian or probabilistic method. Publications have mixed memberships, i.e., each publication belongs to a class with a certain probability (Bleier, 2016). This method better corresponds to the fuzzy nature of scientific communities, the observation that fields and disciplines have cores and peripheries that overlap to some extent (Lietz, 2016).

In this interim report, we use cited references to classify all publications from the *Web of Science* database published in 2010 into 6 sciences, 30 disciplines, and 250 fields. We show that our method clusters publications into meaningful, interpretable, and therefore useful topics at multiple levels of complexity. We evaluate the topics by mapping them to established classes of sciences, disciplines, and fields and show that there are significant dependencies. In addition, we point out how topics from Latent Dirichlet Allocation can be used to delineate publication sets for scientometric studies. The question is not anymore, if an article classification can be delivered. The question is if the error rate is sufficiently low and if the classification can be accepted by customers of scientometric studies.

2 METHODS AND DATA

2.1 UNSUPERVISED AUTOMATED CLUSTERING

Scientometric analyses, whether for evaluation purposes or studying the science system, stand and fall with the used classification system (Glänzel & Schubert, 2003). Essentially, publication classification is a clustering task. The work of Sahrhage et al. (2013) is a recent example of frequentist unsupervised automated clustering. Publications are classified using all available meta data. Despite the failure to cluster large publication sets, it is an important finding that a classifier learning topics from only cited references almost finds the same topics as a hybrid classificatory that learns topics from references and words.

Citation-based clustering classically proceeds by reducing references to those with at least five citations (Van Raan, 1991). Klavans & Boyack (2011) and Boyack & Klavans (2014) proceed in two steps. First, co-citation clusters of cited references are identified as the ideational cores (“intellectual base”) of topics. Second, publications are fractionally assigned to topics based on citations of these cores. By clustering the direct citation graph of all WoS publications from 2001-2010, Waltman & Eck (2012) classify 10.2 million publications. The latter approach comes at the cost of reducing the set of cited references to those indexed in WoS (source items).



FIGURE 1: MIXED-MEMBERSHIP LOGIC OF LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) is a probabilistic approach. The goal is to learn topics based on a set of documents that use terms from a common vocabulary. It rests on the assumptions that a given set of topics exists to which documents are allocated and that each document belongs to each topic with some probability. While documents as mixtures of terms are directly observable, the topics must be learned from the data. Topics are learned by iteratively updating documents as probability distributions over topics and topics as probability distributions over terms (Blei D. , 2012).

LDA meets two requirements from the perspective of modeling socio-cultural systems (Lietz, 2016). First, because documents are probability distributions over topics, the latter have overlapping boundaries. Figure 1 demonstrates the underlying logic for the case of two topics. A document can belong to the red topic to 90% and in the blue topic to 10%. Throughout this report, we will define cores of topics as sets of document that belong to the topic with a minimum probability p . A 0.9-core consists of core documents that define the topic because their probability of belonging to other topics is at most 10%.

The second requirement that is met by LDA is related to topics being probability distributions over terms. A term that belongs to multiple topics also has multiple meanings. The red and blue topics consist of the same terms but with different probabilities (most often zero). Consequently, topics provide different contexts for their terms. Because terms get their meaning from the terms with which they are associated, LDA is a relational method (DiMaggio, Nag, & Blei, 2013).

Originally, LDA was designed for topics to be detected from documents as “bags of words.” Applications of LDA in scientometrics proceed alike by using words (used in titles, abstracts, and as keywords) as terms. Blei & Lafferty (2007) cluster 16,351 publications from the JSTOR collection (1990-1999) which use 19,088 words into 100 topics. Yau et al. (2014) cluster 1,254 publications from WoS which belong to seven known research domains into 50 topics at an average precision of 0.77 and an average recall of 0.67. Suominen & Toivanen (2015) detect 60 topics in a set of 142,656 Finnish publications (1995-2011) from WoS which use 95,664 words, compare topics to OECD classifications, and conclude that there are limits to comparability. Supervised unautomated classifications (expert systems) and unsupervised automated systems (LDA) both have merits that depend on the practical objectives of analysis.

Erosheva, Fienberg, & Lafferty (2004) have applied hybrid of “bilingual” LDA to scientific corpora where topics are detected from word usage and reference citation. They cluster 13,008 publications published in the *Proceedings of the National Academy of Sciences of the United States of America* (1997-2001) and classified in the biological sciences into 8 topics, describe these by high-probability words and references, and find that the traditional discipline classes used by the journal are represented by mixtures of topics. Mund (2014) takes an approach of clustering through words, cited references, and authors. Topics for 3,271 highly-cited publications belonging to 5 disciplines in WoS (2007) are detected, and it is shown that topics that did not exist in the previous year resemble emerging topics.

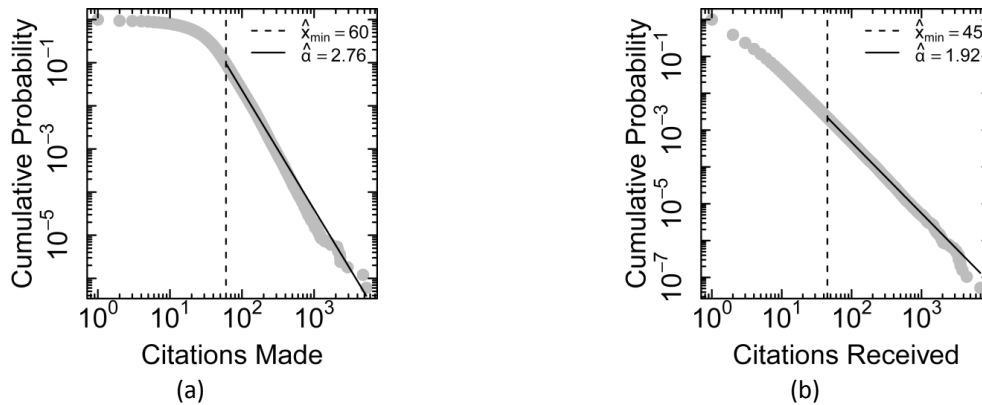


FIGURE 2: REFERENCES PER PAPER AND CITATIONS PER REFERENCE

In this work, relying on previous findings and practices (Klavans & Boyack, 2011; Waltman & Eck, 2012; Sahrhage, Mayland, Maiwald, Gottwald, & Klein, 2013; Boyack & Klavans, 2014), we use cited references for clustering. Other than those approaches, we include non-source references into our analysis, i.e., we model topics based also on the citations of books, non-English works, etc. – in general: all references cited by publications. We employ a conservative procedure to construct reference strings developed by Lietz (2016). Consult the appendix for details.

We employ Stochastic Collapsed Variational Bayesian Inference which allows LDA of big publication sets and vocabularies (Foulds, Boyles, DuBois, Smyth, & Welling, 2013; Bleier, 2016). LDA requires setting the desired number of topics and involves two parameters. The positive continuous parameters α and β set the shape of the probability distributions of topics over terms and of documents over topics. α was learned and β was set according to best practice ($\beta=0.1$). We stopped the LDA inference when the perplexity, a measure for how well the solution in an iteration is predicted by the previous one, did not show major changes anymore.

2.2 DATA

The *Web of Science* (WoS) (Citation Index and Proceedings) covers 1,978,721 distinct publications (identified through UTs) of all document types in the year of study 2010. 1,658,084 publications (84%) cite at least one reference. Figure 2 shows the complementary cumulative distribution functions for citations made by citing publications and citations received by cited references. The fit for the number of citations made (Figure 2a) has an exponent smaller than 3, i.e., the mean and standard deviation are meaningful. A paper has 30 references on average.

These publications cite 19,166,103 distinct references including works not indexed in WoS (non-source items). The fit for the number of citations received (Figure 2b) has an exponent smaller than 2, i.e., there is no typical number of citations. Few citations receive many citations and many receive few.

LDA involves the iterative updating of matrices. These get very large when the number of publications, the number of references, or both are large. Reducing these dimensions is one way to reduce computational costs. Figure 3 shows how the coverage of all 1,658,084 publications is diminished when publications are either required to have specified reference list lengths or when references are required to have a specified number of citations. For example, requiring references to have at least five citations – the historical approach in bibliometrics – reduces the citing set to 1,468,448 publications or 89%. But effects differ for different subsets. While publications in the Citation Index are, like the total, reduced to 89%, publications in the Proceedings set are reduced to 86%. While hardly and reviews get lost, only 6% of the articles but 22% of the letters are removed (Figure 3a).

More importantly, any such filtering introduces an asymmetry or unfairness that derives from the different bibliographic practices of the sciences. The filter of five citations per reference does not remove more than 7%

of all publications in the Agricultural Sciences, Engineering & Technology, the Medical & Health Sciences, and Natural Sciences. But it reduces the Social Sciences to 79% of all publications and practically erases the Humanities (34% coverage). Given that the latter are already underrepresented in the database, filtering renders clustering analyses applicable to the less pluralistic sciences (Figure 3b).

Increasing computer memory is the other way to enable LDA on big datasets.

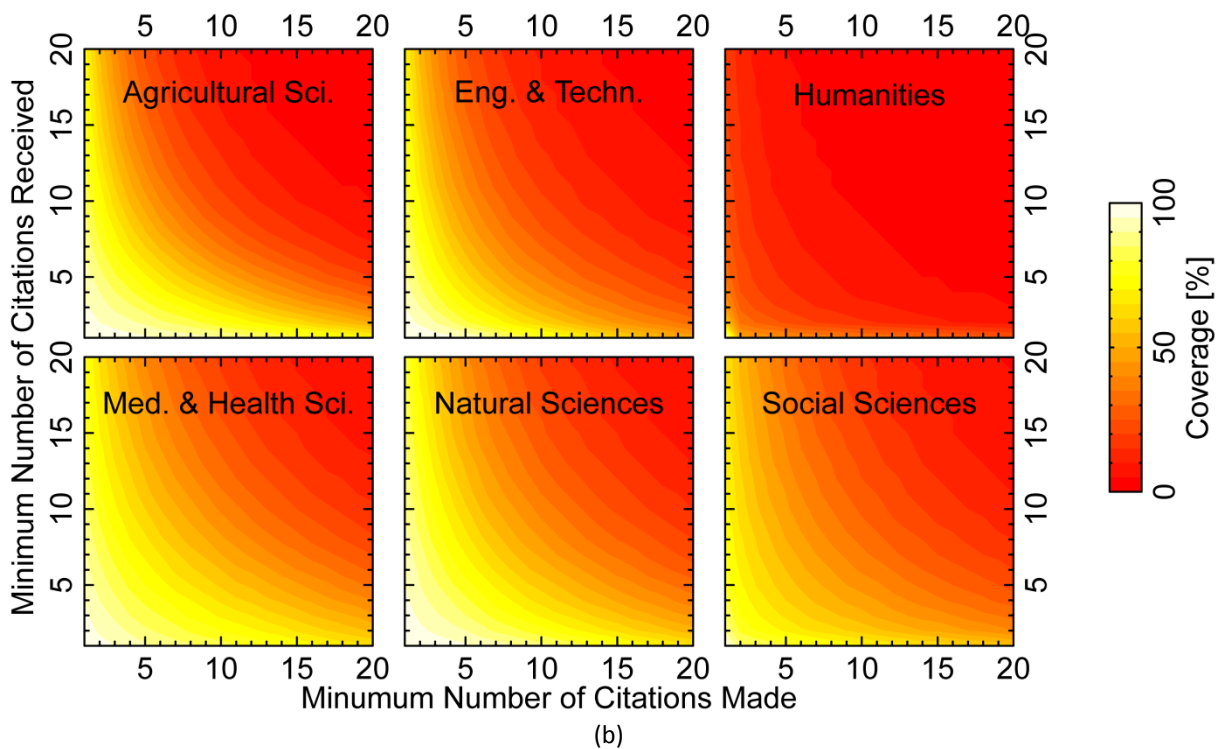
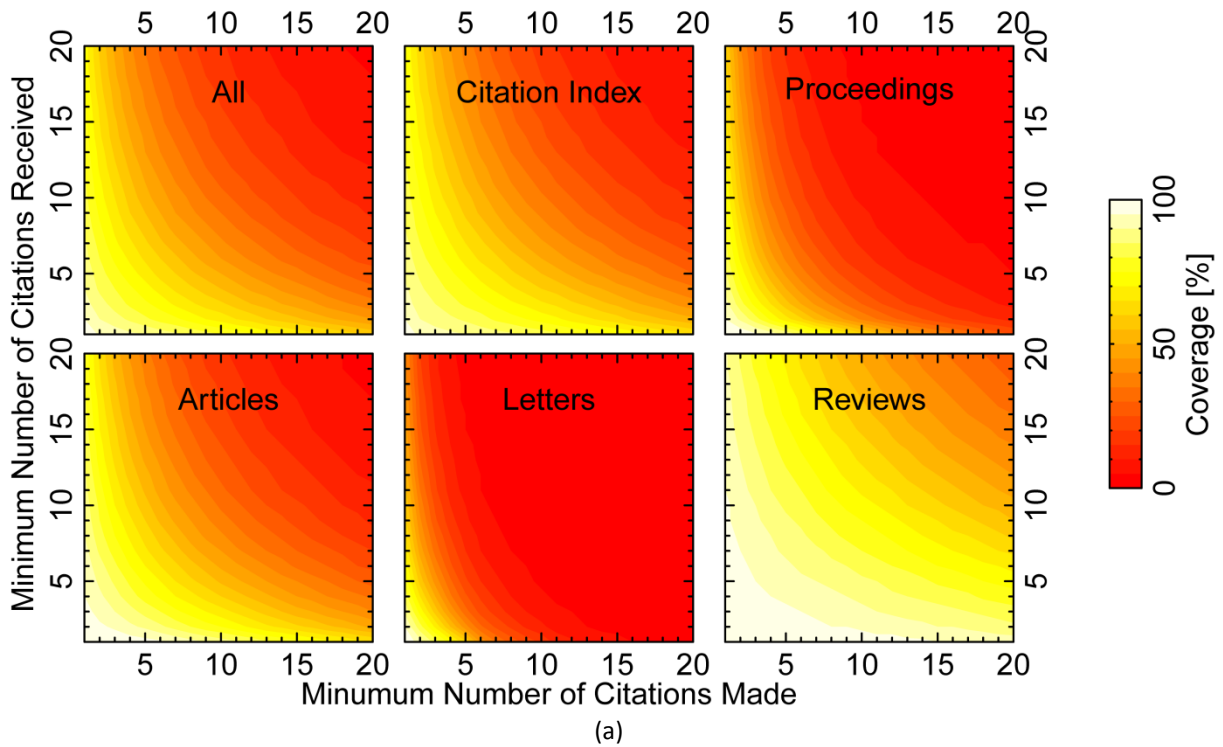


FIGURE 3: LOSS OF COVERAGE WHEN PUBLICATIONS AND REFERENCES ARE FILTERED

3 TOPICS

LDA requires setting the number of topics that should be learned from the data. To arrive at a classification at multiple levels, we take orientation in the classification hierarchy of the Frascati Manual (OECD, 2002) to which a mapping to WoS subject categories exists (Thomson Reuters, 2012). This hierarchy consists of 6 sciences and 39 disciplines mapped to 250 subject categories or fields (excluding “Multidisciplinary Sciences”).

Because references that are only cited once do not contribute to topic modeling, they were removed. This reduced the publication set to 1,571,174 documents. Another 100 publications were held out for technical purposes. Our dataset consists of 1,571,074 citing publications and 7,373,246 cited references. Table 1 reports the number of publications that belong to topic cores at the four probability thresholds shown in Figure 1. For example, only 64% of all publications belong to 0.1-cores of 250 fields.

TABLE 1: REDUCTION OF CORPUS WHEN PROBABILITIES ARE FILTERED

Topics	p≥0.1	p≥0.35	p≥0.65	p≥0.9
6 sciences	1,571,074 (100%)	1,565,621 (100%)	1,265,720 (81%)	789,490 (50%)
39 disciplines	1,568,057 (100%)	1,219,502 (78%)	622,094 (40%)	161,879 (10%)
250 fields	1,006,878 (64%)	457,886 (29%)	193,637 (12%)	28,809 (2%)

Table 6 and Table 7 are descriptions of the resulting topics represented by publications in 0.9-cores. *KeyWords Plus* that stem from terms used in the reference lists of publications (Garfield & Sher, 1993) are mapped to topics. The 30 keywords with highest tf*idf scores are shown.

To interpret these topics and check their plausibility, we draw upon a mapping of topics to WoS/OECD classes. The result is shown in Table 2 and Table 3. Again, the result for the level of 250 topics is given in the supplement. To arrive at these percentages, fractionally counted WoS/OECD classifications are, first, summed over all publications belonging to the 0.9-core of a topic and, second, divided by column or row sums. In each of these tables, the first subtable, the one where columns sum to 100%, tells how much WoS/OECD classes consist of topics. The second subtable tells how much topics consist of WoS/OECD classes. All these mappings show statistically significant dependencies ($p < 0.01$ from chi-squared tests).

The main message is that the Social Sciences and Humanities are largely represented in a single science 2. This is the largest topic, 22% of all 0.9-core publications belong to it. The main keywords of this topic reveal that health is an important subject in this topic. Corroborating this observation, 18% of all publications in the Medical and Health Sciences are in science 2 and 24% of all publications in science 2 are classified in that class. Based on a combined reading of keyword descriptions and class mappings, the labels given in Appendix A are constructed. Instead of one class for the Natural Sciences, there are now three, each with a different disciplinary composition. Science 1 is about Engineering and Technology, science 3 about Biology & Evolution, and science 5 about Chemistry and Climate. The Medical and Health Sciences are hardly mixed with other Sciences, but Oncology now is its own topic. The Agricultural Sciences are basically part of the Natural Sciences and are distributed among sciences 1, 3, and 5.

The 39 disciplines uncovered by LDA are labeled similarly (cf. Table 3 and Table 7). Table 4 allows a comparison to OECD/WoS disciplines. Topics from citation-based clustering more represent actual research problems rather than commonsense notions of scientific disciplines.

TABLE 2: MAPPING OF TOPICS TO WOS/OECD SCIENCES

Topic	Agricultural Sciences	Engineering and Technology	Humanities	Medical and Health Sciences	Natural Sciences	Social Sciences	Total
1	17%	44%	6%	14%	20%	3%	20%
2	4%	11%	61%	18%	9%	86%	22%
3	42%	11%	7%	7%	24%	1%	14%
4	4%	3%	4%	34%	10%	2%	14%
5	29%	19%	16%	3%	25%	5%	16%
6	5%	13%	6%	24%	12%	3%	14%
Total	100%	100%	100%	100%	100%	100%	100%
1	3%	36%	1%	20%	39%	2%	100%
2	1%	8%	9%	24%	16%	43%	100%
3	9%	12%	2%	13%	64%	1%	100%
4	1%	4%	1%	66%	27%	1%	100%
5	6%	19%	3%	6%	63%	3%	100%
6	1%	14%	1%	47%	33%	3%	100%
Total	3%	16%	3%	28%	39%	11%	100%

TABLE 3: MAPPING OF TOPICS TO WOS/OECD DISCIPLINES

Topic	Agriculture, forestry, and fisheries	Animal and dairy science	Art (arts, history of arts, performing arts, music)	Basic medicine	Biological sciences	Chemical engineering	Chemical sciences	Civil engineering	Clinical medicine	Computer and information sciences	Earth and related environmental sciences	Economics and business	Educational sciences	Electrical engineering, electronic engineering	Environmental biotechnology	Environmental engineering	Health sciences	History and archaeology	Industrial Biotechnology	Languages and literature	Law	Materials engineering	Mathematics	Mechanical engineering	Media and communications	Medical engineering	Nano-technology	Other agricultural sciences	Other engineering and technologies	Other humanities	Other natural sciences	Other social sciences	Philosophy, ethics and religion	Physical sciences	Political Science	Psychology	Social and economic geography	Sociology	Veterinary science	Total	
1	0%	0%	3%	1%	1%	0%	0%	0%	1%	1%	9%	0%	0%	0%	0%	0%	0%	13%	0%	8%	0%	0%	7%	0%	0%	0%	0%	0%	4%	2%	0%	0%	6%	0%	0%	1%	1%	2%	2%		
2	0%	0%	0%	1%	1%	0%	1%	1%	2%	1%	0%	0%	1%	1%	3%	2%	0%	0%	0%	0%	0%	0%	0%	13%	0%	2%	0%	1%	2%	0%	1%	0%	0%	0%	0%	0%	0%	0%	2%	1%	
3	1%	0%	22%	0%	0%	0%	0%	1%	1%	2%	0%	0%	26%	2%	3%	0%	1%	44%	0%	4%	66%	0%	0%	0%	0%	0%	3%	2%	33%	0%	61%	24%	0%	0%	88%	1%	7%	37%	0%	3%	
4	0%	0%	0%	1%	2%	0%	0%	1%	1%	0%	33%	0%	0%	2%	2%	0%	1%	50%	0%	0%	1%	0%	1%	1%	1%	1%	1%	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	
5	0%	0%	0%	1%	1%	13%	16%	13%	0%	1%	1%	0%	0%	2%	0%	13%	0%	0%	1%	0%	10%	2%	2%	1%	0%	0%	0%	0%	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	5%	
6	0%	1%	4%	1%	1%	9%	3%	6%	1%	0%	1%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	2%	0%	1%	0%	3%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	3%
7	0%	1%	0%	2%	1%	0%	0%	0%	5%	0%	0%	0%	0%	0%	1%	0%	0%	1%	0%	0%	0%	0%	0%	1%	0%	3%	0%	0%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%
8	0%	0%	0%	1%	0%	0%	0%	3%	3%	1%	0%	9%	0%	0%	0%	1%	1%	1%	3%	0%	1%	6%	2%	12%	0%	1%	3%	0%	0%	1%	0%	1%	0%	10%	4%	0%	0%	1%	1%	3%	
9	0%	0%	2%	1%	1%	2%	0%	4%	2%	24%	1%	11%	0%	8%	1%	3%	3%	1%	0%	0%	0%	6%	6%	3%	1%	0%	1%	8%	4%	1%	0%	1%	1%	1%	0%	2%	1%	5%	2%		
10	3%	1%	0%	1%	1%	1%	1%	1%	0%	2%	1%	6%	0%	1%	3%	1%	2%	0%	0%	0%	0%	0%	2%	2%	0%	2%	0%	32%	11%	4%	1%	0%	0%	0%	0%	1%	1%	1%	1%		
11	1%	3%	0%	0%	9%	1%	1%	1%	4%	2%	0%	0%	2%	4%	11%	2%	2%	2%	2%	0%	1%	3%	2%	6%	0%	1%	0%	6%	2%	3%	1%	0%	1%	0%	0%	0%	0%	4%	1%		
12	0%	0%	0%	1%	0%	1%	7%	11%	0%	1%	0%	0%	0%	3%	1%	0%	0%	1%	0%	0%	1%	6%	0%	0%	11%	2%	2%	0%	2%	0%	0%	5%	0%	0%	0%	0%	0%	0%	0%	2%	
13	1%	0%	2%	4%	2%	3%	0%	1%	0%	2%	5%	3%	0%	1%	1%	4%	0%	1%	0%	0%	1%	0%	1%	10%	0%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	1%	
14	0%	0%	2%	1%	6%	2%	4%	1%	1%	6%	0%	1%	0%	5%	4%	0%	1%	0%	1%	0%	0%	1%	9%	7%	12%	1%	2%	0%	3%	0%	6%	0%	0%	2%	0%	0%	0%	0%	0%	2%	
15	0%	0%	0%	1%	0%	4%	0%	4%	1%	18%	2%	1%	0%	27%	1%	3%	1%	1%	1%	0%	0%	1%	4%	4%	0%	2%	1%	6%	2%	1%	0%	1%	0%	0%	0%	0%	0%	4%	2%		
16	0%	2%	0%	2%	1%	0%	0%	1%	3%	0%	4%	0%	0%	0%	1%	2%	3%	0%	1%	0%	0%	1%	0%	1%	0%	1%	0%	4%	2%	2%	1%	0%	5%	0%	0%	0%	0%	4%	1%	1%	
17	0%	1%	0%	1%	1%	1%	0%	0%	3%	0%	1%	1%	0%	1%	0%	3%	1%	1%	3%	1%	1%	3%	1%	0%	4%	0%	3%	1%	0%	1%	0%	1%	1%	4%	0%	0%	0%	2%	2%		
18	0%	0%	0%	6%	0%	0%	0%	0%	8%	0%	0%	1%	0%	0%	15%	0%	2%	0%	1%	0%	0%	0%	0%	0%	0%	2%	0%	0%	1%	0%	0%	0%	0%	4%	0%	4%	3%	1%	3%		
19	0%	4%	0%	3%	0%	3%	9%	25%	0%	2%	0%	22%	0%	2%	2%	2%	0%	2%	9%	0%	0%	22%	0%	0%	35%	3%	33%	1%	8%	0%	2%	1%	0%	3%	3%	0%	11%	0%	0%	3%	
20	0%	1%	0%	23%	4%	0%	0%	0%	6%	0%	0%	0%	0%	6%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	1%	1%	0%	9%	0%	0%	0%	0%	0%	0%	0%	7%	4%	
21	0%	1%	0%	2%	1%	0%	13%	0%	4%	0%	0%	0%	1%	1%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	2%	1%	0%	0%	0%	0%	0%	0%	0%	6%	3%		
22	0%	0%	1%	0%	0%	0%	0%	3%	2%	3%	7%	0%	0%	4%	0%	1%	1%	2%	0%	0%	0%	1%	15%	3%	0%	0%	3%	0%	2%	0%	3%	0%	1%	12%	0%	1%	0%	0%	0%	3%	
23	0%	0%	4%	1%	0%	0%	0%	13%	0%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	0%	2%	1%	0%	0%	0%	0%	0%	0%	0%	1%	4%		
24	0%	0%	0%	1%	35%	20%	11%	0%	0%	2%	0%	2%	13%	17%	0%	0%	5%	0%	0%	0%	16%	0%	4%	0%	0%	10%	9%	3%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	
25	65%	0%	0%	0%	4%	0%	0%	3%	1%	0%	12%	0%	0%	2%	2%	14%	1%	1%	9%	0%	0%	0%	0%	0%	0%	2%	0%	8%	6%	0%	3%	0%	0%	0%	0%	0%	2%	0%	0%	3%	
26	18%	13%	0%	40%	0%	0%	1%	0%	1%	5%	0%	5%	12%	0%	3%	5%	0%	0%	0%	0%	1%	1%	1%	0%	0%	1%	0%	0%	13%	0%	0%	0%	4%	2%	2%	8%	7%	1%	2%		
27	0%	0%	1%	1%	0%	1%	10%	1%	0%	1%	2%	0%	1%	0%	3%	0%	2%	0%	0%	0%	3%	0%	2%	0%	0%	0%	1%	0%	4%	0%	0%	1%	0%	0%	0%	0%	0%	0%	1%	2%	
28	7%	2%	0%	1%	8%	1%	0%	0%	1%	0%	0%	0%	1%	18%	0%	4%	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	11%	3%	1%	3%	0%	0%	0%	0%	0%	0%	0%	7%	2%		
29	0%	1%	0%	3%	1%	1%	0%	0%	1%	1%	0%	0%	0%	0%	3%	0%	1%	2%	0%	0%	0%	3%	1%	0%	1%	0%	2%	1%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	6%	1%	
30	0%	0%	0%	10%	2%	0%	0%	0%	5%	0%	0%	0%	0%	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	4%	0%	0%	1%	0%	0%	0%	1%	0%	2%	2%		
31	0%	0%	0%	3%	1%	0%	0%	1%	11%	0%	0%	0%	7%	1%	0%	1%	0%	0%	0%	1%	1%	0%	0%	0%	1%	1%	0%	4%	1%	0%	1%	0%	3%	0%	0%	0%	0%	0%	3%		
32	0%	0%	40%	2%	0%	0%	0%	1%	3%	0%	15%	50%	7%	0%	0%	12%	8%	0%	49%	31%	0%	1%	0%	26%	0%	0%	2%	36%	1%	30%	58%	0%	3%	58%	3%	51%	0%	5%			
33	0%	0%	0%	0%	0%	5%	10%	3%	1%	3%	0%	1%	0%	6%	0%	1%	0%	1%	0%	1%	2%	0%	5%	15%	5%	0%	1%	3%	2%	4%	0%	0%	1%	0%	0%	0%	0%	0%	2%		
34	0%	16%	18%	1%	0%	0%	0%	4%	4%	0%	0%	4%	2%	0%	3%	3%	2%	33%	0%	0%	0%	1%	1%	8%	0%	1%	4%	10%	1%	3%	1%	0%	31%	0%	1%	0%	4%	0%			
35	0%	15%	0%	3%	6%	0%	0%	0%	5%	0%	0%	1%	0%	0%	7%	0%	0%	0%	2%	0%	0%	1%	0%	1%	0%	16%	0%	1%	0%	5%	1%	0%	0%	0%	0%	0%	0%	7%	2%		
36	1%	42%	2%	1%	3%	1%	3%	1%	9%	1%	0%	1%	1%	4%	2%	1%	0%	0%	0%	0%	3%	3%	1%	2%	3%	3%	5%	0%	1%	0%	1%	2%	0%	0%	0%	0%	17%	2%			
37	2%	2%	0%	1%	4%	0%	1%	0%	1%	1%	0%	0%	0%	7%	2%	0%	2%	0%	0%	2%	1%	0%	2%	1%	0%	1%	0%	6%	0%	1%	0%	23%	0%	0%	0%	0%	0%	5%			
38	0%	7%	2%	1%	0%	0%	0%	3%	1%	0%	0%	0%	1%	0%	2%	2%	0%	4%	0%	3%	3%	1%	0%	1%	0%	3%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	1%			
39	0%	0%	2%	4%	1%	0%	0%	3%	17%	0%	1%	0%	6%	2%	1%	0%	1%	0%	0%	0%	3%	1%	3%	4%	0%	3%	1%	3%	0%	0%	1%	1%	3%	0%	0%	0%	0%	0%	0%	2%	
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
1	0%	0%	2%	4%	0%	0%	5%	1%	26%	0%	0%	0%	0%	1%	0%	2%	0%	0%	1%	0%	2%	0%	0%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	100%		
2	0%	0%	8%	12%	1%	2%	1%	35%	1%	1%	0%	1%	1%	2%	5%	0%	0%	0%	1%	7%	10%	0%	1%	0%	1%	0%	1%	0%	2%	0%	0%	0%	6%	0%	0%	0%	1%	100%			
3	0%	0%	0%	1%	0%	0%	0%	5%	1%	1%	22%	1%	2%	0%	1%	3%	0%	1%	6%	0%	0%	0%	0%	0%	0%	0%	0%	3%	1%	0%	19%	1%	16%	13%	0%	0%	0%	100%			
4	0%	0%	2%	12%	0%	2%	0%	5%	0%	66%	0%	0%	1%	1%	2%	0%	0%	3%	0%	0%	1%	0%	0%	0%	3%	0%	0%	0%	3%	0%	1%	0%	0%	0%	0%	0%	0%	0%	100%		
5	0%	0%	0%	0%	0%	2%	33%	2%	0%	1%	0%	1%	0%	1%	0%	0%	0%	0%	8%	1%	0%	0%	8%	1%	0%	1%	0%	1%	0%	0%	45%	0%	0%	0%	0%	0%	0%	100%			
6	0%	0%	5%	4%	4%	19%	3%	8%	0%	4%	0%	1%	0%	5%	17%	0%	0%	0%	0%	0%	0%	14%	1%	2%	0%	2%	2%	0%	1%	0%	1%	0%	0%	7%	0%	0%	0%	1%	100%		
7	0%	0%	11%	13%	0%	4%	0%	58%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	4%	1%	0%	0%	0%	1%	0%	0%	2%	0%	0%	5%	5%	0%	0%	0%	0%	0%	0%	100%		
8	0%	0%	2%	0%	0%	1%	1%	19%	0%	0%																															

TABLE 4: OVERVIEW OF OECD/WOS CLASSES AND DISCIPLINES FROM TOPIC MODELING

OECD/WoS Science	OECD/WoS Discipline	LDA Discipline
1: Natural Sciences	1: Mathematics 2: Computer and information sciences 3: Physical sciences and astronomy 4: Chemical sciences 5: Earth and related environmental sciences 6: Biological sciences 7: Other natural sciences	1: Cosmology 2: Medicine I 3: Social Sciences 4: Geoscience 5: Materials Science 6: Materials and Health Science 7: Oncology I 8: Particle Physics 9: Computer Science 10: Medicine II 11: Biofuels 12: Nanoscience I 13: Marine Science & Neurology 14: Molecular Biology 15: Robotics & Artificial Intelligence 16: Clinical Medicine & Pharmacology 17: Physics & Medicine 18: Mental Health 19: Nanoscience II 20: Cell Medicine 21: Molecular Medicine 22: Quantum Mechanics 23: Health Science & String Theory 24: Chemical Engineering 25: Environmental Science 26: Evolutionary Science 27: Chemistry & Complexes 28: Environmental Biology 29: Basic Medicine 30: Neurological Diseases 31: Oncology II 32: Psychology & Education 33: Materials & Complexes 34: Cognition & Language 35: Cell Biology 36: Environment & Aerosols 37: Astronomy 38: Medicine III 39: Bioinformatics
2: Engineering and Technology	1: Civil engineering 2: Electrical eng, electronic eng 3: Mechanical engineering 4: Chemical engineering 5: Materials engineering 6: Medical engineering 7: Environmental engineering 8: Environmental biotechnology 9: Industrial biotechnology 10: Nano-technology 11: Other engineering and technologies	
3: Medical and Health Sciences	1: Basic medical research 2: Clinical medicine 3: Health sciences	
4: Agricultural Sciences	1: Agriculture, forestry, fisheries 2: Animal and dairy science 3: Veterinary science 4: Other agricultural science	
5: Social Sciences	1: Psychology 2: Economics and business 3: Educational sciences 4: Sociology 5: Law 6: Political science 7: Social and economic geography 8: Media and communication 9: Other social sciences	
6: Humanities	1: History and archaeology 2: Languages and literature 3: Philosophy, ethics and religion 4: Art 5: Other Humanities	

4 APPLICATIONS AND DISCUSSION

The topics we have detected allow top-down and bottom-up publication retrieval. Top-down means that it is known which topics must be selected to delineate a publication set. The question scientometricians then need no answer is if the certainty that a publication belongs to a topic must be high or can be allowed to be low. As we see in Table 1, sets are small when probabilities are high.

Bottom-down means that, from a large number of fine-grained topics, an unknown topics or set of topics is identified for which publications are then retrieved. This is a classical field delineation task (Mayr, Scharnhorst, Larsen, Schaer, & Mutschke, 2014). To delineate a field using LDA, consider the example of Social Network Science. Lietz (2016) has evaluated a set of 106 publications, 58 of which belong to this research domain. Of these, 43 belong to one or more of the 250 topics identified by LDA. Field 53 contains 21 and field 39 contains 15 of those 43 publications. These two fields are described in Table 5. Both contain major descriptors of Social Network Science (Lietz, 2016). On the contrary, when the 48 publications are considered that do not belong to Social Network Science, no particular topics are revealed and fields 53 and 39 are not singled out as sources of false positives.

TABLE 5: DESCRIPTION OF SELECTED FIELDS

53: Complex Networks
COMPLEX NETWORKS, DYNAMICS, SCALE-FREE NETWORKS, EVOLVING NETWORKS, SOLID LUBRICANTS, MASS-SPECTROMETER, CORONARY-ARTERY, MODULARITY, B-CELL LYMPHOMA, METABOLIC NETWORKS, SYNCHRONIZATION, MODEL, ULTRAMETRICITY, PHYSICS, TOPOLOGY, SMALL-WORLD NETWORKS, INTERNET, CRITICAL-BEHAVIOR, COMMUNITY STRUCTURE, AWARENESS ROUTING STRATEGY, SYSTEMS, CASCADING FAILURES, COMPLEX HETEROGENEOUS NETWORKS, MARKETS, SCALE-FREE, RESOLUTION, ORGANIZATION, OPTIMIZATION, HIERARCHICAL ORGANIZATION, OUTBREAKS
39: Business & Management
ABSORPTIVE-CAPACITY, COMPETITIVE ADVANTAGE, INNOVATION, RESOURCE-BASED VIEW, STRATEGIC ALLIANCES, FIRM, RESEARCH-AND-DEVELOPMENT, FIRM PERFORMANCE, CAPABILITIES, KNOWLEDGE, ULTRASTRONG MAGNETIC FIELDS, STRATEGY, MARKET ORIENTATION, PERFORMANCE, JOINT VENTURES, PRODUCT DEVELOPMENT, INDUSTRY, FIRMS, DYNAMIC CAPABILITIES, KNOWLEDGE TRANSFER, DEMENTIA, TECHNOLOGY, PERSPECTIVE, ORGANIZATIONS, MULTINATIONAL-CORPORATIONS, SOCIAL-STRUCTURE, ALLIANCES, ALLIANCE FORMATION, SR-90, BUSINESS PERFORMANCE

Both on the level of 39 disciplines and 250 fields, irregularities have been found that need further clarification. For example, discipline 13 is a mix of Marine Science & Neurology, discipline 23 of Health Science & String Theory, and field 39 is described by “ULTRASTRONG MAGNETIC FIELDS”. This may be a result from stopping the inference process to early and will be checked.

In sum, LDA has been shown to uncover meaningful, interpretable, and therefore useful topics at multiple levels of complexity. These topics are often at the intersection of classical, expert-based classifications of science. Furthermore, these topics may have been overlooked by experts. LDA is really a lense that allows seeing patterns in large amounts of data.

The big vocabulary of more that 7 million terms has turned out to seriously increase computational costs. It took the detection of 6 topics 3 hours, 39 topics 12 hours, and 250 topics 72 hours to converge. In these processes, 6 GB, 18 GB, and 60 GB computer memory, respectively, were required. The detection of 1,600 topics will take our 500 GB machine to the limit. In the course of our analyses, the Java implementation of topic modeling (Matte) turned out to be unable to process our dataset.

To deliver what we have announced, we will detect topics for WoS publications from 1990 onwards. Despite the fast inference implemented, it is already clear that this will not be possible with the parameter settings we have used for this study. We intend to keep the publications in the Humanities and Social Sciences that we would otherwise lose due to reference filtering by classifying publications through citations and word usage. To allow fine-grained field delineation, we will try to detect $6 \cdot 6.5^3 = 1,648$ and $6 \cdot 6.5^4 = 10,710$ topics. The 6, 39, and 250 topics we have detected so far do not form a hierarchical classification system in the sense that the levels are detected independently. For a real hierarchy, future projects may resort to hierarchical LDA (Blei, Griffiths, Jordan, & Tenenbaum, 2004).

APPENDIX A: DESCRIPTION OF TOPICS

TABLE 6: DESCRIPTION OF 6 SCIENCES

1: Natural Sciences (Engineering and Technology)
SYSTEMS, DESIGN, ALGORITHM, NETWORKS, SYSTEM, OPTIMIZATION, PERFORMANCE, MODEL, ALGORITHMS, STABILITY, FLOW, NANOPARTICLES, CAPACITY, INFECTION, GENERATION, CHANNELS, TRANSMISSION, MANAGEMENT, WIRELESS NETWORKS, FABRICATION, LIGHT, DIVERSITY, LASER, PARTICLES, SPECTROSCOPY, NANOCRYSTALS, DIAGNOSIS, CHILDREN, EQUATIONS, DYNAMICS
2: Social Sciences & Humanities
PERFORMANCE, MODEL, CARE, HEALTH, BEHAVIOR, EDUCATION, MANAGEMENT, IMPACT, CHILDREN, STUDENTS, KNOWLEDGE, UNITED-STATES, GROWTH, POLICY, WOMEN, ATTITUDES, RISK, POLITICS, GENDER, PERSPECTIVE, INFORMATION, WORK, QUALITY, DEPRESSION, SCIENCE, ADOLESCENTS, SYSTEMS, PERCEPTIONS, OUTCOMES, PREVALENCE
3: Natural Sciences (Biology & Evolution)
EVOLUTION, IDENTIFICATION, ESCHERICHIA-COLI, PLANTS, EXPRESSION, INFECTION, DNA, GROWTH, GENE, RESISTANCE, DIVERSITY, BACTERIA, STRAINS, ARABIDOPSIS-THALIANA, GENES, SEQUENCE, GENE-EXPRESSION, GENOME, PHYLOGENY, INFECTIONS, SEQUENCES, PROTEIN, SACCHAROMYCES-CEREVISIAE, MODEL, POPULATIONS, REMOVAL, ARABIDOPSIS, DISEASE, STARS, SYSTEM
4: Medical and Health Sciences (Oncology)
EXPRESSION, CANCER, ACTIVATION, CELLS, IN-VIVO, IN-VITRO, GENE-EXPRESSION, GENE, CARCINOMA, MICE, SURVIVAL, APOPTOSIS, DIFFERENTIATION, BRAIN, PROTEIN, BREAST-CANCER, THERAPY, ALZHEIMERS-DISEASE, DISEASE, CHEMOTHERAPY, OXIDATIVE STRESS, GROWTH, MUTATIONS, TUMORS, PROLIFERATION, INHIBITION, STEM-CELLS, PATHWAY, CENTRAL-NERVOUS-SYSTEM, PHOSPHORYLATION
5: Natural Sciences (Chemistry & Climate)
MODEL, COMPLEXES, BEHAVIOR, DERIVATIVES, MECHANICAL-PROPERTIES, DYNAMICS, VARIABILITY, TEMPERATURE, GROWTH, CHEMISTRY, CRYSTAL-STRUCTURE, MICROSTRUCTURE, LIGANDS, SYSTEM, WATER, SYSTEMS, CLIMATE-CHANGE, CLIMATE, DESIGN, MANAGEMENT, PATTERNS, VEGETATION, COMPOSITES, TRANSPORT, CARBON, CONSERVATION, CRYSTAL-STRUCTURES, MODELS, PRECIPITATION, DEFORMATION
6: Medical and Health Sciences
RISK, DISEASE, MANAGEMENT, MORTALITY, THERAPY, MODEL, TRIAL, MYOCARDIAL-INFARCTION, DIAGNOSIS, RISK-FACTORS, SURGERY, PREVALENCE, OUTCOMES, CHILDREN, CARDIOVASCULAR-DISEASE, WATER, ASSOCIATION, BEHAVIOR, SYSTEM, MODELS, COMPLICATIONS, BLOOD-PRESSURE, PERFORMANCE, FOLLOW-UP, ACUTE MYOCARDIAL-INFARCTION, METAANALYSIS, POPULATION, CORONARY-HEART-DISEASE, SYSTEMS, DOUBLE-BLIND

TABLE 7: DESCRIPTION OF 39 DISCIPLINES

1: Cosmology
UNIVERSE, DARK ENERGY, GRAVITY, COSMOLOGICAL CONSTANT, SUPERNOVAE, GENERAL-RELATIVITY, COSMOLOGY, CONSTRAINTS, INFLATION, CONSTANT, PROBE WMAP OBSERVATIONS, MODELS, BLACK-HOLES, FIELD, SPACE, HAWKING RADIATION, ACCELERATING UNIVERSE, QUINTESSENCE, ENERGY, QUANTUM-GRAVITY, PERTURBATIONS, RECORD, MODEL, EVOLUTION, HUBBLE-SPACE-TELESCOPE, EQUATION-OF-STATE, GEOMETRY, MATTER, THERMODYNAMICS, FIELD-THEORY
2: Medicine I
LARGE-EDDY SIMULATION, MODEL, SHOULDER, FOLLOW-UP, MANAGEMENT, FLOWS, ISOTROPIC TURBULENCE, TURBULENCE, DEATH, COMBUSTION, CHANNEL FLOW, ACTIVATION, REPAIR, EQUATIONS, TEARS, NAVIER-STOKES EQUATIONS, FLOW, DIAGNOSIS, RESECTION, CANCER, DISLOCATION, LONG-QT SYNDROME, EXPRESSION, FEMOROACETABULAR IMPINGEMENT, TUMORS, LES, DYNAMICS, MAGNETOHYDRODYNAMIC TURBULENCE, SUDDEN CARDIAC DEATH, STABILIZATION
3: Social Sciences
POLITICS, DEMOCRACY, POLICY, STATE, GOVERNANCE, MODEL, INSTITUTIONS, LAW, IMPACT, UNITED-STATES, GROWTH, EUROPE, GLOBALIZATION, GOVERNMENT, SATISFACTION, POWER, PARTICIPATION, COUNTRIES, CITY, ECONOMY, POVERTY, QUALITY, WORLD, DETERMINANTS, TRUST, INEQUALITY, BEHAVIOR, LOYALTY, MIGRATION, MODELS
4: Geoscience
EVOLUTION, TRACE-ELEMENT, TECTONIC EVOLUTION, CONTINENTAL-CRUST, ROCKS, CONSTRAINTS, DEFORMATION, VOLCANIC-ROCKS, U-PB, GEOCHEMISTRY, GEOCHRONOLOGY, MANTLE, TECTONIC IMPLICATIONS, UPPER-MANTLE, MAGMATISM, IN-VITRO, U-PB GEOCHRONOLOGY, PETROGENESIS, TECTONICS, CALIFORNIA, SUBDUCTION, LITHOSPHERE, BENEATH, SCAFFOLDS, LITHOSPHERIC MANTLE, ZIRCON, ZONE, ORIGIN, BELT, COMPLEX
5: Materials Science
AB-INITIO, DENSITY-FUNCTIONAL THEORY, ELECTRONIC-STRUCTURE, MOLECULES, AUGMENTED-WAVE METHOD, SEMICONDUCTORS, ENERGY, TOTAL-ENERGY CALCULATIONS, SPECTROSCOPY, DENSITY, GAUSSIAN-BASIS SETS, ADSORPTION, ATOMS, EXCHANGE, METALS, GENERALIZED GRADIENT APPROXIMATION, SURFACE, DYNAMICS, TRANSITION, THIN-FILMS, FILMS, BASIS-SETS, OXIDES, MAGNETORESISTANCE, SYSTEMS, SPECTRA, TEMPERATURE, APPROXIMATION, WAVE-FUNCTIONS, SURFACES
6: Materials and Health Science
TITANIUM-DIOXIDE, DEGRADATION, WATER, TIO2, PERFORMANCE, AMORPHOUS-ALLOYS, FILMS, OXIDATION, VISIBLE-LIGHT, STRENGTH, ANATASE, PHOTOCATALYTIC ACTIVITY, VISIBLE-LIGHT IRRADIATION, THIN-FILMS, SENSITIZED SOLAR-CELLS, EFFICIENCY, NANOPARTICLES, EXERCISE, IRRADIATION, MECHANICAL-PROPERTIES, PHOTOCATALYTIC DEGRADATION, DECOMPOSITION, HUMAN SKELETAL-MUSCLE, FABRICATION, PHOTODEGRADATION, ALLOYS, OXIDE, FORMING ABILITY, POWDERS, SKELETAL-MUSCLE
7: Oncology I

ENDOTHELIAL GROWTH-FACTOR, ANGIOGENESIS, EXPRESSION, CANCER, INTERFERON-ALPHA, IN-VIVO, WETTABILITY, POSITRON-EMISSION-TOMOGRAPHY, TUMOR ANGIOGENESIS, THERAPY, SURVIVAL, CELLS, VEGF, NEOVASCULARIZATION, BREAST-CANCER, SUNITINIB, METASTASIS, RENAL-CELL CARCINOMA, BEVACIZUMAB, CARCINOMA, ENDOTHELIAL-CELLS, GROWTH-FACTOR, TUMOR-GROWTH, SYSTEM, TUMORS, GROWTH, PHASE-III TRIAL, SUPERHYDROPHOBIC SURFACES, NEUROENDOCRINE TUMORS, MICE
8: Particle Physics
PHYSICS, STANDARD MODEL, MODEL, DECAYS, HADRON COLLIDERS, NUCLEI, QCD, SCATTERING, DECAY, PERCUTANEOUS CORONARY INTERVENTION, MASSES, COLLISIONS, SEARCH, ENERGIES, MASS, SEVERE PLASTIC-DEFORMATION, DARK-MATTER, LHC, BREAKING, ACUTE MYOCARDIAL-INFARCTION, DETECTOR, SUPERSYMMETRY, STATES, DEFORMATION, MATTER, CROSS-SECTIONS, BEHAVIOR, PHOTOPRODUCTION, BOSON, CORPORATE GOVERNANCE
9: Computer Science
OPTIMIZATION, ALGORITHM, SYSTEMS, MODEL, GENETIC ALGORITHM, MODELS, DESIGN, UNCERTAINTY, ALGORITHMS, SYSTEM, STRAINS, EMERGENCE, GLOBAL OPTIMIZATION, TABU SEARCH, MANAGEMENT, SEARCH, INFECTIONS, PANTON-VALENTINE LEUKOCIDIN, GENETIC ALGORITHMS, METHICILLIN-RESISTANT, EVOLUTION, UNITED-STATES, RADICAL PROSTATECTOMY, COLONIZATION, FRAMEWORK, EVOLUTIONARY GAMES, IMPACT, CRITERIA, PROGRAMS, OUTCOMES
10: Medicine II
FLAVONOIDS, EXTRACTS, PHENOLIC-COMPOUNDS, POLYPHENOLS, CAPACITY, RISK, RETURNS, ANTIOXIDANT ACTIVITY, ANTHOCYANINS, VEGETABLES, ASSAY, CRITICALLY-ILL PATIENTS, FRUITS, PLUS RIBAVIRIN, SEPTIC SHOCK, L., IN-VITRO, INTENSIVE-CARE-UNIT, GENOME-WIDE ASSOCIATION, IDENTIFICATION, INFECTION, COMBINATION THERAPY, VOLATILITY, ACUTE-RENAL-FAILURE, ASSOCIATION, PEGINTERFERON ALPHA-2A, STOCK RETURNS, THERAPY, MEDICINAL-PLANTS, FRUIT
11: Biofuels
TRANSESTERIFICATION, SOYBEAN OIL, VEGETABLE-OILS, FUEL, BIODIESEL PRODUCTION, BROMINATED FLAME RETARDANTS, ESTERS, DIESEL-ENGINE, RAPESEED OIL, POLYBROMINATED DIPHENYL ETHERS, POLYCHLORINATED-BIPHENYLS, BEHAVIOR, PERFORMANCE, WASTE COOKING OIL, FREE FATTY-ACIDS, METHANOL, METHYL-ESTERS, OIL, BIODIESEL, SEED OIL, FUEL PRODUCTION, BLENDS, PBDES, PALM OIL, ESTERIFICATION, COMBUSTION, OPTIMIZATION, SUNFLOWER OIL, VEGETABLE-OIL, PERSISTENT ORGANIC POLLUTANTS
12: Nanoscience I
GRAPHITE, FILMS, COMPLEX NETWORKS, CARBON NANOTUBES, EPITAXIAL GRAPHENE, SHEETS, TRANSPORT, GAS, BERRYS PHASE, TRANSISTORS, COMPOSITES, LARGE-AREA, FIELD-EFFECT TRANSISTORS, RAMAN-SPECTROSCOPY, BILAYER GRAPHENE, GRAPHENE, CHEMICAL-VAPOR-DEPOSITION, THIN-FILMS, DYNAMICS, FUNCTIONALIZATION, POLARIZATION, ALDEHYDES, GROWTH, NANOCOMPOSITES, TRANSPARENT, NANOSHEETS, SUSPENDED GRAPHENE, GRAPHITE OXIDE, CARBON, SCATTERING
13: Marine Science & Neurology
PHYTOPLANKTON, SYNAPTIC PLASTICITY, LONG-TERM POTENTIATION, COASTAL WATERS, COINTEGRATION, MODEL, NANOPARTICLES, THERMAL-CONDUCTIVITY, HIPPOCAMPAL-NEURONS, OCEAN, WATER, GLUTAMATE RECEPTORS, UNIT-ROOT TESTS, DYNAMICS, CARBON, MARINE-PHYTOPLANKTON, UNIT-ROOT, SYNAPSES, TIME-SERIES, FLOW, ZOOPLANKTON, NUCLEUS-ACCUMBENS, RAT HIPPOCAMPUS, FLUIDS, MARINE-SEDIMENTS, NEURONS, TRANSMITTER RELEASE, SYNAPTIC-TRANSMISSION, COMMUNITY STRUCTURE, PLASTICITY
14: Molecular Biology
MOLECULAR-DYNAMICS, MOLECULAR-DYNAMICS SIMULATIONS, DYNAMICS, CRYSTAL-STRUCTURE, MODEL, SIMULATIONS, SYSTEMS, STABILIZATION, STABILITY, ESCHERICHIA-COLI, SIMULATION, BINDING, PROTEINS, PROTEIN, FORCE-FIELD, ROBUST STABILITY, MECHANISM, DIFFUSION, DESIGN, TRANSPORT, FREE-ENERGY, LIQUID WATER, WATER, H-INFINITY CONTROL, EQUATION, STRUCTURE PREDICTION, CRITERIA, SPECTROSCOPY, RESOLUTION, RECOGNITION
15: Robotics & Artificial Intelligence
FEATURES, SYSTEMS, TRACKING, STABILITY, NETWORKS, STABILIZATION, ALGORITHM, SCALE, AGENTS, MULTIAGENT SYSTEMS, RECOGNITION, ALGORITHMS, SLAM, COORDINATION, COOPERATIVE CONTROL, SIMULTANEOUS LOCALIZATION, IMAGES, MOBILE ROBOTS, MOTION, CONSENSUS, MODEL, GRAPH CUTS, DESIGN, OBJECT DETECTION, IMAGE, MODELS, FLOCKING, SEGMENTATION, REGISTRATION, VISION
16: Clinical Medicine & Pharmacology
D DEFICIENCY, FLUCONAZOLE, AMPHOTERICIN-B, D SUPPLEMENTATION, INFECTION, EPIDEMIOLOGY, RISK, VORICONAZOLE, FUNGAL-INFECTIONS, SERUM 25-HYDROXYVITAMIN-D, D INSUFFICIENCY, MS, LIPOSOMAL AMPHOTERICIN-B, HYPOVITAMINOSIS-D, THERAPY, PREVALENCE, DISABILITY, MANAGEMENT, PLACEBO-CONTROLLED TRIAL, CHILDREN, DISEASE, PRECIPITATION, BLOOD-STREAM INFECTIONS, 25-HYDROXYVITAMIN-D, MULTIPLE-SCLEROSIS, DOUBLE-BLIND, DIAGNOSTIC-CRITERIA, DIAGNOSIS, GONADOTROPIN-RELEASING-HORMONE, INVASIVE CANDIDIASIS
17: Physics & Medicine
SUPERCONDUCTIVITY, BA0.6K0.4FE2AS2, LAYERED QUATERNARY COMPOUND, LAYERED SUPERCONDUCTOR, 43 K, SUPERCONDUCTORS, COMPOUND, GAPS, STATE, CHRONIC KIDNEY-DISEASE, PHASE-DIAGRAM, KLEBSIELLA-PNEUMONIAE, SERUM CREATININE, GLOMERULAR-FILTRATION-RATE, OUTCOMES, MORTALITY, LAO1-XFXFEAS, ENTEROBACTERIACEAE, FESE, RISK-FACTORS, RESISTANCE, THERAPY, ACUTE ISCHEMIC-STROKE, PULMONARY TUBERCULOSIS, EMERGENCE, MYCOBACTERIUM-TUBERCULOSIS, IRON, EPIDEMIOLOGY, ESCHERICHIA-COLI, BOSE-EINSTEIN CONDENSATION
18: Mental Health
PREVALENCE, DEPRESSION, MAJOR DEPRESSION, SYMPTOMS, RANDOMIZED CONTROLLED-TRIAL, POSTTRAUMATIC-STRESS-DISORDER, PRIMARY-CARE, QUALITY-OF-LIFE, DISORDERS, SCALE, VALIDITY, FOLLOW-UP, RISK-FACTORS, WOMEN, RELIABILITY, DISORDER, VALIDATION, METAANALYSIS, POPULATION, COGNITIVE-BEHAVIORAL THERAPY, ANXIETY, HEALTH, LOW-BACK-PAIN, UNITED-STATES, PSYCHIATRIC-DISORDERS, COMORBIDITY, DOUBLE-BLIND, QUESTIONNAIRE, PSYCHOMETRIC PROPERTIES, DISABILITY
19: Nanoscience II
NANOPARTICLES, NANOCRYSTALS, OPTICAL-PROPERTIES, GROWTH, NANORODS, NANOSTRUCTURES, NANOWIRES, PARTICLES, QUANTUM DOTS, FABRICATION, SIZE, ARRAYS, GOLD NANOPARTICLES, PHOTOLUMINESCENCE, INNOVATION, FILMS, THIN-FILMS, SPECTROSCOPY, PERFORMANCE, ENHANCED RAMAN-SCATTERING, SEMICONDUCTOR NANOCRYSTALS, SILVER NANOPARTICLES, ROUTE, METAL NANOPARTICLES, SURFACE, KNOWLEDGE, FIRMS, LUMINESCENCE, EMISSION, ZINC-OXIDE

20: Cell Medicine
DENDRITIC CELLS, IN-VIVO, T-CELLS, EXPRESSION, REGULATORY T-CELLS, CUTTING EDGE, ACTIVATION, IMMUNE-RESPONSES, RESPONSES, MICE, LYMPHOCYTES, IFN-GAMMA, RECEPTOR, TGF-BETA, ANTIGEN, INDUCTION, IMMUNITY, INFECTION, INFLAMMATORY-BOWEL-DISEASE, NATURAL-KILLER-CELLS, INFLAMMATION, DIFFERENTIATION, TOLERANCE, TOLL-LIKE RECEPTORS, INTERFERON-GAMMA, INNATE IMMUNITY, NF-KAPPA-B, PERIPHERAL-BLOOD, RHEUMATOID-ARTHRITIS, DISEASE
21: Molecular Medicine
CROSS-COUPLING REACTIONS, ORGANIC-SYNTHESIS, UNITED-STATES, COMPLEXES, DERIVATIVES, CHILDREN, INFECTION, ARYL HALIDES, BOND FORMATION, ARYLBORONIC ACIDS, HALIDES, OBSTRUCTIVE PULMONARY-DISEASE, ALKYNES, COUPLING REACTIONS, VIRUS, EFFICIENT, STEREOSELECTIVE-SYNTHESIS, FUNCTIONALIZATION, ARYL CHLORIDES, ROOM-TEMPERATURE, OVERWEIGHT, CYCLIZATION, DIRECT ARYLATION, A H1N1, OBESITY, HETEROCYCLES, ARYLATION, PALLADIUM, RANDOMIZED CONTROLLED-TRIAL, ADOLESCENTS
22: Quantum Mechanics
ENTANGLEMENT, STATES, QUANTUM, COMPUTATION, STATE, SYSTEMS, DECOHERENCE, CRYPTOGRAPHY, CAVITY, QUANTUM COMPUTATION, LIGHT, TELEPORTATION, DYNAMICS, QUBITS, PHOTONS, SUDDEN-DEATH, MODEL, FIELD, INFORMATION, GENERATION, MECHANICS, SEPARABILITY, COMMUNICATION, ATOMS, SYSTEM, VARIABILITY, SPIN, INTERFERENCE, NOISE, MOTION
23: Health Science & String Theory
CARDIOVASCULAR-DISEASE, CORONARY-HEART-DISEASE, INSULIN-RESISTANCE, METABOLIC SYNDROME, RISK, BLOOD-PRESSURE, MYOCARDIAL-INFARCTION, MORTALITY, RISK-FACTORS, ATHEROSCLEROSIS, C-REACTIVE PROTEIN, OBESITY, DISEASE, CORONARY-ARTERY-DISEASE, WOMEN, STRING THEORY, PREVALENCE, MEN, ASSOCIATION, RANDOMIZED CONTROLLED-TRIAL, HEART-DISEASE, POPULATION, HYPERTENSION, SUPERGRAVITY, MELLITUS, CHOLESTEROL, QCD, DENSITY-LIPOPROTEIN CHOLESTEROL, DIABETES-MELLITUS, PREVENTION
24: Chemical Engineering
DEVICES, FIELD-EFFECT TRANSISTORS, EFFICIENCY, POLYMERS, THIN-FILM TRANSISTORS, CONJUGATED POLYMERS, MORPHOLOGY, LIGHT-EMITTING-DIODES, PERFORMANCE, COPOLYMERS, BLOCK-COPOLYMERS, PHOTOVOLTAIC CELLS, ADSORPTION, DIODES, POLYMER, WATER, REMOVAL, SORPTION, FILMS, TRANSPORT, BLENDS, AQUEOUS-SOLUTIONS, MOBILITY, POLYMER SOLAR-CELLS, AQUEOUS-SOLUTION, DERIVATIVES, EQUILIBRIUM, TRANSFER RADICAL POLYMERIZATION, CHARGE-TRANSPORT, SOLAR-CELLS
25: Environmental Science
NITROGEN, VEGETATION, ORGANIC-MATTER, DYNAMICS, CARBON, CLIMATE-CHANGE, SOIL, FOREST, GROWTH, MANAGEMENT, MODEL, MICROBIAL BIOMASS, ECOSYSTEMS, YIELD, LEAF-AREA INDEX, CARBON-DIOXIDE, WATER, TEMPERATURE, VARIABILITY, BIOMASS, SEQUESTRATION, CLIMATE, USE EFFICIENCY, BOREAL FOREST, LAND-USE, PRODUCTIVITY, PATTERNS, TILLAGE, DECOMPOSITION, RESPIRATION
26: Evolutionary Science
EVOLUTION, DIVERSITY, CONSERVATION, POPULATIONS, BIODIVERSITY, PATTERNS, MITOCHONDRIAL-DNA, SEQUENCES, POPULATION-STRUCTURE, DNA, ECOLOGY, SELECTION, BIRDS, PHYLOGENY, SEXUAL SELECTION, DISPERSAL, IDENTIFICATION, INFERENCE, COMMUNITIES, SPECIATION, SPECIES RICHNESS, BODY-SIZE, CLIMATE-CHANGE, PHYLOGEOGRAPHY, NATURAL-SELECTION, MARKERS, MOLECULAR PHYLOGENY, POPULATION, MULTILOCUS GENOTYPE DATA, SIZE
27: Chemistry & Complexes
BENZOFURAN, COMPLEXES, CRYSTAL-STRUCTURE, DERIVATIVES, LIGANDS, X-RAY, METAL-COMPLEXES, ACID, DIFFRACTION, MERIT, CRYSTAL-STRUCTURES, CHEMISTRY, SOLID-PHASE EXTRACTION, BINDING, COPPER(II) COMPLEXES, LIGAND, AGENTS, PERSONAL CARE PRODUCTS, WASTE-WATER, REACTIVITY, TANDEM MASS-SPECTROMETRY, LUMINESCENCE, SEWAGE-TREATMENT PLANTS, PERFORMANCE, STRUCTURAL-CHARACTERIZATION, OLEFIN POLYMERIZATION, THERMOELECTRIC PROPERTIES, SOLVATE, CRYSTAL, MOLECULAR-STRUCTURE
28: Environmental Biology
ARABIDOPSIS-THALIANA, PLANTS, EXPRESSION, GENE-EXPRESSION, ESCHERICHIA-COLI, ARABIDOPSIS, IDENTIFICATION, PARITY-CHECK CODES, ABSICIS-ACID, GENE, BIOSYNTHESIS, PROTEIN, GENES, THALIANA, TOLERANCE, SIGNAL-TRANSDUCTION, GROWTH, STRESS, OXIDATIVE STRESS, INFECTION, RESISTANCE, BACTERIA, HYDROGEN-PEROXIDE, SALICYLIC-ACID, VIRULENCE, SALT TOLERANCE, STRAINS, METABOLISM, DISEASE RESISTANCE, PSEUDOMONAS-AERUGINOSA
29: Basic Medicine
P-GLYCOPROTEIN, TOTAL MESORECTAL EXCISION, EXPRESSION, PHARMACOKINETICS, IN-VITRO, IDENTIFICATION, METABOLISM, PREOPERATIVE RADIOTHERAPY, HUMAN LIVER-MICROSOMES, CARCINOMA, MULTIDRUG-RESISTANCE, DIAGNOSIS, PREGNANE-X-RECEPTOR, BLOOD-BRAIN-BARRIER, HUMAN LIVER, POSTOPERATIVE CHEMORADIOTHERAPY, INFECTION, THERAPY, RANDOMIZED-TRIAL, CHEMOTHERAPY, PURIFICATION, LOCAL RECURRENCE, DRUG-DRUG INTERACTIONS, 2ND-GRADE FLUID, SALT EXPORT PUMP, DRUG-INTERACTIONS, POLYMERASE-CHAIN-REACTION, HOMOTOPY ANALYSIS METHOD, CONSTITUTIVE ANDROSTANE RECEPTOR, EXISTENCE
30: Neurological Diseases
ALZHEIMERS-DISEASE, DEMENTIA, MILD COGNITIVE IMPAIRMENT, PARKINSONS-DISEASE, A-BETA, BRAIN, AMYOTROPHIC-LATERAL-SCLEROSIS, TRANSGENIC MICE, PROTEIN, DIAGNOSIS, LEWY BODIES, IN-VIVO, AMYLOID PRECURSOR PROTEIN, DISEASE, MOUSE MODEL, PATHOLOGY, ALPHA-SYNUCLEIN, FRONTOTEMPORAL LOBAR DEGENERATION, IMPAIRMENT, CEREBROSPINAL-FLUID, CREUTZFELDT-JAKOB-DISEASE, COGNITIVE IMPAIRMENT, PRECURSOR PROTEIN, PROGRESSION, OXIDATIVE STRESS, AGGREGATION, APOLIPOPROTEIN-E, PREVALENCE, GENE, IN-VITRO
31: Oncology II
CHEMOTHERAPY, CARCINOMA, SURVIVAL, THERAPY, TRIAL, PHASE-III TRIAL, REFRACTION, CLOAK, GEFITINIB, GROWTH-FACTOR RECEPTOR, METAMATERIALS, RANDOMIZED-TRIAL, CELL LUNG-CANCER, CANCER, CISPLATIN, ADJUVANT CHEMOTHERAPY, ADENOCARCINOMA, NEGATIVE REFRACTION, PHASE-II, PACLITAXEL, FREQUENCIES, PHASE-II TRIAL, BREAST-CANCER, LIGHT, DOCETAXEL, RADIATION-THERAPY, EXPRESSION, INDEX, RADIOTHERAPY, ERLOTINIB
32: Psychology & Education
EDUCATION, STUDENTS, BEHAVIOR, PERFORMANCE, PERSONALITY, MODEL, MOTIVATION, SELF, PERCEPTIONS, ATTITUDES, WORK, KNOWLEDGE, GENDER, ACHIEVEMENT, METAANALYSIS, PERSPECTIVE, IDENTITY, SCHOOL, CHILDREN, ADJUSTMENT, VALIDITY, SATISFACTION, CLASSROOM, ADOLESCENTS, SELF-ESTEEM, CONSEQUENCES, VALIDATION, BELIEFS, PSYCHOLOGY, OUTCOMES

33: Materials & Complexes
METAL-ORGANIC FRAMEWORKS, COORDINATION POLYMERS, COMPLEXES, CRYSTAL-STRUCTURES, MAGNETIC-PROPERTIES, HYDROTHERMAL SYNTHESIS, CRYSTAL-STRUCTURE, DESIGN, BUILDING-BLOCKS, CHEMISTRY, LIGANDS, HYDROGEN STORAGE, METAL-ORGANIC FRAMEWORK, NETWORK, NETWORKS, LIGAND, FRAMEWORKS, CLUSTERS, SIGNAL RECOVERY, ADSORPTION, ARCHITECTURES, SEPARATION, RECONSTRUCTION, COORDINATION POLYMER, SOLIDS, FRAMEWORK, ACID, POLYMERS, CHAINS, ALGORITHM
34: Cognition & Language
PERCEPTION, CORTEX, BRAIN, FMRI, HUMAN BRAIN, ATTENTION, WORKING-MEMORY, PREFRONTAL CORTEX, RECOGNITION, INFORMATION, MEMORY, REPRESENTATION, TASK, LANGUAGE, ACTIVATION, EVENT-RELATED POTENTIALS, TRANSCRANIAL MAGNETIC STIMULATION, SHORT-TERM-MEMORY, FUNCTIONAL MRI, CHILDREN, RECOGNITION MEMORY, PARIETAL CORTEX, EVENT-RELATED FMRI, BRAIN POTENTIALS, ACQUISITION, HUMANS, DISCRIMINATION, VISUAL-CORTEX, EYE-MOVEMENTS, REPRESENTATIONS
35: Cell Biology
DIFFERENTIATION, EXPRESSION, IN-VITRO, SELF-RENEWAL, PROGENITOR CELLS, MOUSE, STEM-CELLS, EMBRYONIC STEM-CELLS, TRANSPLANTATION, GENERATION, DEFINED FACTORS, HOSPITAL CARDIAC-ARREST, BONE-MARROW, LINES, IMPLANTATION, MICE, INDUCTION, HUMAN SOMATIC-CELLS, GENE-EXPRESSION, FIBROBLASTS, SURVIVAL, HUMAN FIBROBLASTS, GENE, STROMAL CELLS, CARDIOPULMONARY-RESUSCITATION, HUMAN BLASTOCYSTS, IN-VIVO, PLURIPOTENCY, CULTURE, THROMBOSIS
36: Environment & Aerosols
PARTICULATE MATTER, PARTICLES, EMISSIONS, SOURCE APPORTIONMENT, TRANSPORT, ATMOSPHERE, VOLATILE ORGANIC-COMPOUNDS, CHEMICAL-COMPOSITION, SECONDARY ORGANIC AEROSOL, BLACK CARBON, AIR-POLLUTION, CHEMISTRY, AEROSOL, OZONE, UNITED-STATES, ATMOSPHERIC AEROSOLS, OPTICAL COHERENCE TOMOGRAPHY, CHIP, URBAN, OPTICAL-PROPERTIES, PM2.5, MODEL, SYSTEM, POSITIVE MATRIX FACTORIZATION, EXISTENCE, POLYCYCLIC AROMATIC-HYDROCARBONS, SYSTEMS, POLLUTION, SIZE DISTRIBUTIONS, ORBITS
37: Astronomy
ACTIVE GALACTIC NUCLEI, EVOLUTION, DIGITAL SKY SURVEY, STAR-FORMATION, EMISSION, STARS, CATALOG, GALAXIES, LUMINOSITY FUNCTION, STAR-FORMING GALAXIES, MILKY-WAY, ELLIPTIC GALAXIES, SPIRAL GALAXIES, X-RAY, XMM-NEWTON, HIGH-REDSHIFT, INITIAL MASS FUNCTION, DISCOVERY, HUBBLE-SPACE-TELESCOPE, ACCRETION, MASS, MOLECULAR CLOUDS, GAS, DATA RELEASE, EXPRESSION, LARGE-MAGELLANIC-CLOUD, SPECTROSCOPY, INTERSTELLAR-MEDIUM, PHOTOMETRY, TELESCOPE
38: Medicine III
DOUBLE-BLIND, RHEUMATOID-ARTHRITIS, EFFICACY, THERAPY, INFlixIMAB, MANAGEMENT, TRIAL, METHOTREXATE, ASYMPTOTIC METHODS, BOUNDARY-VALUE-PROBLEMS, VARIATIONAL ITERATION METHOD, DISEASE-ACTIVITY, FOLLOW-UP, PLACEBO-CONTROLLED TRIAL, HOMOTOPY PERTURBATION METHOD, DISEASE, ETANERCEPT, TRANSITIONAL-CELL CARCINOMA, MODIFYING ANTIRHEUMATIC DRUGS, SAFETY, LIPID-PEROXIDATION, PORCINE MODEL, PROGRESSION, RANDOMIZED CONTROLLED-TRIAL, ADALIMUMAB, CEREBRAL ANEURYSMS, RANDOMIZED-TRIAL, ANKYLOSING-SPONDYLITIS, EXPERIENCE, IDENTIFICATION
39: Bioinformatics
CLASSIFICATION, SACCHAROMYCES-CEREVISIAE, CLASSIFIERS, CATHETER ABLATION, SUPPORT VECTOR MACHINES, HOMOLOGOUS RECOMBINATION, ALGORITHM, DOUBLE-STRAND BREAKS, CANCER, ALGORITHMS, S-PHASE, NEURAL-NETWORKS, MAMMALIAN-CELLS, PULMONARY VEIN ISOLATION, CHROMATIN, REGRESSION, MACHINES, CELL-CYCLE, COMPLEX, DNA METHYLATION, REPAIR, RECOGNITION, GENE-EXPRESSION, MUTATIONS, METHYLATION, FOLLOW-UP, DIAGNOSIS, MANAGEMENT, GENE, EXPRESSION

APPENDIX B: REFERENCE DATA

Tables: wosci_p_2015_17.cited_references and wosci_p_2015_17.cited_references.

Data subsets are constructed using Query 1.

QUERY 1: REFERENCES CITED IN 2010 (WOSCI)

```
CREATE TABLE wosci_pubs_2010 AS
SELECT pk_source_item_records, ut
FROM wosci_p_2015_17.source_item_records
JOIN wosci_p_2015_17.source_issue_records ON pk_source_issue_records =
fk_source_issue_records
WHERE py = 2010;

CREATE TABLE wosci_cited_references_2010 AS
SELECT cited_references.*
FROM wosci_p_2015_17.cited_references
JOIN wosci_pubs_2010 ON pk_source_item_records = fk_source_item_records;
```

Cited references are coded into three types of facts (articles, books, and chapters) using the fields given in Table 8 and the procedure described in Table 9 (Query 2). Examples are given in Table 10. Using cited volumes and pages is considered the best approach to distinguish books/chapters and articles (Moed, 2005, pp. 122-3). Our procedure corrects for books or proceedings series that use volume numbers. Proceedings articles can be found in the article and chapter categories, depending on if a volume is cited.

TABLE 8: FIELDS USED FOR CODING CITED REFERENCES

Field	Description
ca	Cited author
cy	Cited reference year
cw	Cited work
cv	Cited volume
cp	Cited page
rs_artn	Cited article number
rs_doi	Cited DOI

TABLE 9: CODING OF FACTS

```
IF volume cited
  IF page OR article number OR doi cited THEN CODE AS 'article'
  ELSE CODE AS 'book'
ELSE
  IF page OR article number OR doi cited THEN CODE AS 'chapter'
  ELSE CODE AS 'book'
```

QUERY 2: CODING OF FACTS (WOSCI)

```
CREATE TABLE wosci_facts_2010 AS
SELECT
  pk_cited_references,
  CASE WHEN cv IS NULL THEN (
    CASE WHEN cp IS NULL THEN (
      CASE WHEN rs_artn IS NULL THEN (
```

```

CASE WHEN rs_doi IS NULL THEN 'BOOK' ELSE 'CHAPTER' END
) ELSE 'CHAPTER' END
) ELSE 'CHAPTER' END
) ELSE (
CASE WHEN cp IS NULL THEN (
CASE WHEN rs_artn IS NULL THEN (
CASE WHEN rs_doi IS NULL THEN 'BOOK' ELSE 'ARTICLE' END
) ELSE 'ARTICLE' END
) ELSE 'ARTICLE' END
) END AS subtype,
UPPER (SUBSTR(REGEXP_SUBSTR(ca, '[^,]+', 1, 1), 1,
8))||'_'||cy||'_'||CASE WHEN cv IS NULL THEN (
CASE WHEN cp IS NULL THEN (
CASE WHEN rs_artn IS NULL THEN (
CASE WHEN rs_doi IS NULL THEN UPPER(REGEXP_SUBSTR(cw, '[^ ]+', 1,
1)) ELSE UPPER(REGEXP_SUBSTR(cw, '[^ ]+', 1, 1)) END
) ELSE UPPER(REGEXP_SUBSTR(cw, '[^ ]+', 1, 1)) END
) ELSE UPPER(REGEXP_SUBSTR(cw, '[^ ]+', 1, 1)) END
) ELSE (
CASE WHEN cp IS NULL THEN (
CASE WHEN rs_artn IS NULL THEN (
CASE WHEN rs_doi IS NULL THEN UPPER(REGEXP_SUBSTR(cw, '[^ ]+',
1, 1)) ELSE UPPER(SUBSTR(cw, 1, 1)) END
) ELSE UPPER(SUBSTR(cw, 1, 1)) END
) ELSE UPPER(SUBSTR(cw, 1, 1)) END
) END||CASE WHEN cv IS NULL THEN (
CASE WHEN cp IS NULL THEN (
CASE WHEN rs_artn IS NULL THEN (
CASE WHEN rs_doi IS NULL THEN ' ' ELSE ' '||UPPER(rs_doi) END
) ELSE ' '||UPPER(SUBSTR(rs_artn, 2)) END -- substring necessary
to cut off blanks at beginning of rs_artn
) ELSE ' '||UPPER(cp) END
) ELSE (
CASE WHEN cp IS NULL THEN (
CASE WHEN rs_artn IS NULL THEN (
CASE WHEN rs_doi IS NULL THEN ' ' ELSE ' '||UPPER(rs_doi) END
) ELSE ' '||UPPER(SUBSTR(rs_artn, 2)) END -- substring necessary
to cut off blanks at beginning of rs_artn
) ELSE ' '||UPPER(cp) END
) END AS fact
FROM wosci cited references 2010

```

TABLE 10: EXAMPLES OF CODED FACTS

pk_cited_refere nces	ca	cy	cw	cv	cp	rs_art n	rs_doi	pubtyp e	fact
2099077896	CARR, LD	2009	NEW J PHYS	11	55049	55049	10.1088/1367-2630/11/5/055049	ARTICLE	CARR_2009_N_55049
2049786897	GEFEN, D	2005	COMMUNICATI ONS ASSI	16	91	5	(null)	ARTICLE	GEFEN_2005_C_91
1897633388	MOLITORIS , BA	2007	NAT CLIN PRACT NEPHR	3	439	(null)	10.1038/ncpneph0551	ARTICLE	MOLITORI_2007_N_439
1897633390	MUKHERJE E, E	2007	CLIN NEPHROL	68	186	(null)	(null)	ARTICLE	MUKHERJE_2007_C_186
2131881368	ROBERTS, I	2000	COCHRANE DB SYST REV	2	(null)	(null)	(null)	BOOK	ROBERTS_2000_COCHRANE
1897633417	TACCONI, FS	2009	CRIT CARE	13	(null)	R15	10.1186/cc7713	ARTICLE	TACCONI_2009_C_R15
1954223152	JONES, S	2004	GENOME BIOL	5	(null)	226	(null)	ARTICLE	JONES_2004_G_226
1883128353	WANG, ZP	2005	J GEOPHYS RES	110	(null)	(null)	10.1029/2004J0005548	ARTICLE	WANG_2005_J_10.1029/2004J0005548
2137769543	LI, M	2010	MATH PROBL ENG	(null)	97454	15726 4	10.1155/2010/157264	CHAPTER	LI_2010_MATH_97454
2172778514	BLOSCHL, G	2005	ENCY HYDROLOGICAL SC	(null)	135	9	(null)	CHAPTER	BLOSCHL_2005_ENCY_135
2237620830	MCCAFFER TY, E	1971	DISCUSS FARADAY SOC	(null)	239	(null)	10.1039/DF9715200239	CHAPTER	MCCAFFER_1971_DISCUSS_239

2269904030	LOFAS, S	1990	J CHEM SOC CHEM 1101	(null)	1526	(null)	(null)	CHAP TER	LOFAS_1990_J_1526
2199770867	CHO, JY	2008	MEDIAT INFLAMM	(null)	(null)	298010	10.1155/2008/298010	CHAP TER	CHO_2008_MEDIAT_298010
2099078067	SJOSTRAN D, T	2004	J HIGH ENERGY PHYS	(null)	(null)	53	(null)	CHAP TER	SJOSTRAN_2004_J_053
2240998660	WALKER, LA	2009	J MOL CELL CARDIOL	(null)	(null)	(null)	10.1016/J.YJMCC.2309.09.010	CHAP TER	WALKER_2009_J_10.1016/J.YJMCC.2309.09.010
2269904031	MASCINI, M	2009	APTAMERS BIOANALYSIS	(null)	(null)	(null)	(null)	BOOK	MASCINI_2009_APTAMERS

Table 11 reports the numbers of cited references, numbers of facts, and compression rates for the citation index and the proceedings. The most citation variants for a unique fact are exist in the citation index for the article category.

TABLE 11: COMPRESSION OF CITED REFERENCES INTO FACTS

		Cited references	Facts	Compression
Citation index	Article	38,495,641	13,058,272	66%
	Book	5,440,392	3,074,577	43%
	Chapter	2,433,773	1,792,035	26%
	Total	46,369,806	17,908,905	61%
Proceedings	Article	2,799,481	1,916,601	32%
	Book	1,112,884	792,685	29%
	Chapter	489,125	411,011	16%
	Total	4,401,490	3,115,847	29%

The next step is to merge the citation index and the proceedings and identify the distinct citations of facts by papers. The latter are now represented by uts (Query 3).

QUERY 3: MERGING CITATION INDEX AND PROCEEDINGS

```
CREATE TABLE citation_2010 AS
SELECT DISTINCT *
FROM (
  SELECT ut, fact
  FROM wosci_pubs_2010
  JOIN wosci_p_2015_17.cited_references ON pk_source_item_records =
fk_source_item_records
  JOIN wosci_facts_2010 ON cited_references.pk_cited_references =
wosci_facts_2010.pk_cited_references
  UNION
  SELECT ut, fact
  FROM wospr_pubs_2010
  JOIN wospr_p_2015_17.cited_references ON pk_source_item_records =
fk_source_item_records
  JOIN wospr_facts_2010 ON cited_references.pk_cited_references =
wospr_facts_2010.pk_cited_references
)
```

5 LITERATURE

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* , 55 (4), pp. 77-84.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics* , 1 (1), pp. 17-35.
- Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, & B. Schölkopf, *Advances in Neural Information Processing Systems* (pp. 17-24). Cambridge: MIT Press.
- Blei, D. (2012). Probabilistic topic models. *Communications of the ACM* , 55 (4), pp. 77-84.
- Bleier, A. (2016). *Approximate Methods for Bayesian Models with Applications to the Social Sciences*. Dissertation. Leipzig: Universität Leipzig.
- Boyack, K., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology* , 65 (4), pp. 670-685.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* , 41 (6), pp. 570-606.
- Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America* , 101, pp. 5220-5227.
- Foulds, J., Boyles, L., DuBois, C., Smyth, P., & Welling, M. (2013). Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* , pp. 446-454.
- Garfield, E., & Sher, I. H. (1993). KeyWords Plus™—algorithmic derivative indexing. *Journal of the American Society for Information Science* , 44 (5), pp. 298-299.
- Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics* , 56 (3), pp. 357-367.
- Klavans, R., & Boyack, K. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology* , 62 (1), pp. 1-18.
- Lietz, H. (2016). *Scale-Free Identity: The Emergence of Social Network Science*. Dissertation. Duisburg: Universität Duisburg-Essen.
- Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). Bibliometric-enhanced information retrieval. *Lecture Notes in Computer Science* , 8416, pp. 798-801.
- Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer.
- Mund, C. (2014). *Identification of Emerging Scientific Topics in Bibliometric Datasets*. Dissertation. Karlsruhe: Karlsruher Institut für Technologie.
- OECD. (2002). *Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development*. Paris: OECD Publishing.

Sahrhage, E., Mayland, J., Maiwald, G., Gottwald, S., & Klein, T. (2013). *Klassifikation und Ähnlichkeitsanalyse von mathematischen Publikationen unter Einbeziehung aller verfügbaren Daten*. Projektbericht. Bielefeld: Kompetenzzentrum Bibliometrie.

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology* .

Thomson Reuters. (2012). *OECD Category Scheme*. Retrieved from <http://ipscience-help.thomsonreuters.com/inCites2Live/researchAreaSchema/oeCdCategoryScheme/version/2>

Van Raan, A. (1991). Fractal geometry of information space represented by co-citation clustering. *Scientometrics* , 20 (3), pp. 439-449.

Waltman, L., & Eck, N. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology* , 63 (12), pp. 2378-2392.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics* , 100 (3), pp. 767-786.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management* , 6, pp. 1513-1531.