NJC

PAPER

View Article Online
View Journal | View Issue

Cite this: *New J. Chem.*, 2015, 39, 8807

Chemical bibliographic databases: the influence of term indexing policies on topic searches†

Gilles Niel,*[a] Fabrice Boyrie[b] and David Virieux[a]

Received (in Montpellier, France) 30th April 2015,
Accepted 28th August 2015

DOI: 10.1039/c5nj01077b

www.rsc.org/njc

A comparative study of the three main chemical information systems (Scifinder, Web of Science and Scopus) was performed by studying the indexing policies of titles, abstracts and keywords within selected literature articles. Various chemical expressions were introduced as topic searches to illustrate the different search tools related to term indexing. The resulting article lists were compared two-by-two by means of a script designed to identify common reference lists and specific ones to each editor. Analyzing these specific reference lists reveals that only partial coverage areas of references should be expected when querying a single platform. The discussion covers the term and keyword indexing policies, their influence on the retrievability of references and on the retrievability of the highly cited papers.

1. Introduction

If many previous studies compare bibliographic databases[1–3] in terms of citation analysis very few ones deal with the herein concerned topic. Falagas compared the strengths and weaknesses of PubMed, Scopus, WoS and Google Scholar providing an interesting overview about their main available search tools.[4] This author introduced a single keyword as a topic search but did not provide any hit counts resulting from this particular search. An other in-depth analysis on chemical databases was proposed by Zass and shows some inconsistencies in indexing policies of the Chemical Abstracts Service (CAS) but his results were not compared with other major bibliographic platforms.[5] These preliminary studies prompted us to analyze the consequences of term indexing policies on the number and on the consistency of retrieved answers by comparing the three above-mentioned platforms.

Term indexing has received much attention for many years from the herein compared information systems. The CAS indexes journal articles, among other document types, since the beginning of the twentieth century in a highly hierarchical way. The bibliographic CAplus database contains currently more than 40 million records covering a wide range of chemical domains including biochemistry, organic, macromolecular and applied chemistry as well as inorganic, analytical and physical chemistry.[6] The CAS's title coverage comes close to 10 000 titles among them 1700 key journals are gathered to form a core journal list.[7] From the outset CAS's indexing policy is document-oriented by the CAS that provides indexed terms from titles, abstracts and author keywords to a large extent in the CAplus database. In its current version, supplementary information using a hierarchical set of controlled terms is also provided.[8] At the top level of the hierarchy a reference is first associated with one of the 80 CAS's sections and then indexing is divided into three main categories: concepts, substance related information and supplementary indexing terms.[9] The concept category contains one or several subject headings at the first level and then terms or text-modifying phrases at the second level, both levels constituting the controlled vocabulary. Supplementary terms are keywords added by the editor that may be either different from controlled terms or may be excerpted from author keywords. The substance related information is categorized in a similar way *i.e.* the first level displays substance identifiers such as the Registry Number, the common chemical names linked with the official chemical name. The second level consists of index terms excerpted from the controlled vocabulary and also from CAS's specific terms such as substance roles.[10] Thus this powerful indexing relies on both CAplus and Registry[11] databases enabling the user to retrieve a large reference set while using a text-only querying language.[12] Moreover Scifinder, the CAS's web interface, enables reference searching from both CAplus and MEDLINE[13] databases, the indexing of the latter relying on the National Library of Medicine's controlled vocabulary thesaurus, named Medical Subject Headings (MeSH).[14]

[a] *Institut Charles Gerhardt Montpellier (ICGM), UMR 5253 CNRS-UM-ENSCM, Ecole Nationale Supérieure de Chimie, 8, rue de l'Ecole Normale, 34296 Montpellier, France. E-mail: gilles.niel@enscm.fr*
[b] *Institut Charles Gerhardt Montpellier (ICGM), UMR 5253 CNRS-UM-ENSCM, Université de Montpellier, Place E. Bataillon, 34095 Montpellier, France*

† Electronic supplementary information (ESI) available: Additional information on the diversity of studied domains (Table S1), on the expanded timespan from 1990–2005 (Table S2), the studied references in Table 8 (Doc 1), the studied references in Table 10 (Doc 2) and the source code of the script (Doc 3). See DOI: 10.1039/c5nj01077b

Among the whole WoS's databases, the Science Citation Index Expanded™ (SCIE) gives access to more than 40 million records from a large range of scientific domains.[15] The 8500 indexed journals cover a larger set of scientific domains divided into 182 categories related to mathematics, physics, chemistry, biology, medicine, engineering, *etc.*[16] Besides the title, abstract and author keyword fields, WoS provides ESI,† gathered in the Keywords Plus® field.[17] This information results from an algorithmic process that excerpts terms appearing at least two times in the titles of the cited references of a processed article.[18]

SciVerse Scopus[19] indexes more than 21 000 titles in all scientific topics classified into four domains: social sciences, physics, life and health sciences.[20] These two latter domains are especially well represented and the total record number comes close to 50 millions today. The term indexing policy includes titles, abstracts, author keywords as well as matched terms. These matched terms include chemical names, CAS Registry Numbers, trade names, manufacturer names and index keywords. These index keywords form the hierarchically controlled vocabulary gathered in several thesauri such as the Compendex index,[21] EMTREE index,[22] MeSH, Species index, and GeoBase subject index.[23] This list is non-exhaustive but refers to the main indexes concerned by this comparative study.

A second important factor concerns the query language and the related query tools. The CAS introduced gradually the use of a natural language to process queries by developing a computed generation of index entries from natural language phrases.[24–26] In recent years this led to the natural language query (NLQ) system, an algorithmic process that breaks down phrases into concepts.[27] Different instructions of the process were first described by J. Williams[28] then thoroughly analyzed by A. Ben Wagner.[29] The last step of the algorithmic process consists in truncating any remaining term that is not parsed in a prior instruction, thus the term 'organocatalysis' will furnish references containing the terms: organocatalysis, organocatalyst(s), organocatalytic, and organocatalys(z)ed. The main characteristics of the NLQ system lie in avoiding: (i) the use of Boolean operators that are interpreted like prepositions,[30] (ii) the use of proximity operators, and (iii) any knowledge about specific field searches. Prepositions are only used to break down phrases into simpler concepts. The NLQ process enables the end-user to focus on the scientific content owing to an easy-to-use topic search interface that may appear simpler at the outset by comparison to those of WoS or of Scopus. Both latter editors provide either basic or advanced search modes that enable searches on specific fields. WoS and Scopus provide a more classic use of Boolean operators including proximity operators thus giving the searcher a higher precision on the queried expressions. In advanced query mode, many different search fields of WoS and Scopus are searchable using a quite simple syntax based on field codes.

To assess the influence of some factors such as term indexing and journal coverage, we selected some single terms or short expressions that attempt to be representatives of different chemical domains such as organic and inorganic chemistry, analytical and physical chemistry, chemistry related to energy and fuels or materials science, biochemistry and molecular biology, and biotechnology and biochemical research methods (see ESI,† Table S1). All selected terms and expressions were submitted to the query interfaces of Scifinder, WoS and Scopus and the resulting hit sets were thoroughly analyzed.

2. Results

2.1. Querying methods

This study was limited to some document types such as journal articles, book chapters, conference papers, notes, letters and reviews because all these citation types cover the most informative part of the chemical literature. As a second argument, chemists frequently need to refer to experimental procedures that are more often embedded in journal articles than in other document types. Thus meeting abstracts, errata and corrections were discarded from the initial queries. Patents were also discarded herein because they would require a parallel study owing to their intrinsic indexing that is distinct from the ones of academic papers. The CAplus and Medline databases were queried through the Explore References by Research Topic of Scifinder. The Science Citation Index Expanded and Conference Proceedings Citation Index of WoS were selected while querying these databases in advanced search mode. The three subject areas – life sciences, health sciences and physical sciences – were queried from Scopus's databases such as Embase and Medline.[31] Most queries were performed in 2010 but some queries were performed in previous years to check the reproducibility of the initial results over a larger timespan. Only lists of English-written papers were saved and then exported into a standard bibliographic format for comparison.

Table 1 displays the Scifinder's specific queries corresponding to some selected terms and expressions and then the whole filtering process towards the selection of unique articles. Thus column 2 displays the queried terms as they were typed in the Research Topic form of Scifinder's interface and column 3 specifies which candidate list was chosen at the next step unless otherwise noted. Filtering by year and language leads to crude hit counts (column 4). Column 5 displays the hit counts after combining answer sets when required. The citation column 6 refers to all citations after automatic removal of duplicates from the CAplus and Medline databases while column 7 corresponds to article counts after the selection of document types such as journal articles, book chapters, conference papers, notes, letters and reviews. In some entries, discarding patents from this study involves a dramatic decrease between the citation column and the article one. Other document types, *i.e.* meeting abstracts, errata and corrections, were discarded from citation lists by means of a script, named Iddup, that will be described below. Unique articles in column 8 result from parsing each reference list by this script so that each list does not contain any duplicate reference. The differences between the reference counts of columns 7 and 8 result from incomplete duplicate removal between the CAPlus and Medline and from some errata that could not be filtered during the document type selection.

Table 1 Scifinder's queried expressions and the filtering process

Entry	Queried expression	Candidate	Hit counts[a]	Combine answer sets[b]	Citations[c]	Articles[d]	Unique articles[e]
1	Allene	The concept ''allene''	366	917	771	630	585
2	Allenes	The concept ''allenes''	879				
3	Organocatalysis	The concept ''organocatalysis''	1034	—	827	667	603
4	Peptidomimetics	The concept ''peptidomimetics''	726	—	610	314	305
5	Agostic interactions	The concept ''agostic interactions''	75	—	59	53	51
6	Battery electrodes	The concept ''battery electrodes''	1554	—	1493	807	806
7	Graphene biosensors	The concept ''graphene biosensors''	119	—	95	89	87
8	N-Heterocyclic carbene	''N-Heterocyclic carbene'' as entered	668	793	671	508	450
9	N-Heterocyclic carbenes	''N-Heterocyclic carbenes'' as entered	231				
10	Modified nucleoside	The concept ''modified nucleoside''	199	213	153	106	98
11	Modified nucleosides	The concept ''modified nucleosides''	213				
12	Phosphine ligand	''Phosphine ligand'' as entered	190	378	330	274	253
13	Phosphine ligands	''Phosphine ligands'' as entered	225				
14	Renewable feedstock	The concept ''renewable feedstock''	211	—	187	113	113
15	Copper (Cu) catalyzed arylation	See text	76	—	65	53	51
16	Hybrid materials and nanoparticles	See text	266	—	217	179	177
17	Viscosity of ionic liquids	The two concepts ''viscosity'' and ''ionic liquids'' were present anywhere in the reference	509	—	429	343	335
18	Band gap in solar cells	The two concepts ''band gap'' and ''solar cells'' closely associated with one another	554	—	526	431	430
19	Statistical analyses of DNA microarrays	References were found where the two concepts ''statistical analyses'' and ''DNA microarrays'' were present anywhere in the reference	154	—	150	98	93
20	Surface area in mesoporous materials	The two concepts ''surface area'' and ''mesoporous materials'' closely associated with one another	372	—	346	296	294

[a] Crude hit counts after filtering by year and language. [b] After combining answer sets when required. [c] After automatic removal of duplicates from the CAplus and Medline databases. [d] Selection of some document types. [e] After parsing by the Iddup script.

As pointed out by Ben Wagner, singular form *vs.* plural form queries in Scifinder may lead to somewhat different results. Therefore we tested each term or expression under both forms.[29] In most cases the hit counts are equal except for entries 1, 2, 8, 9, 10, 11, 12 and 13. For example the answer lists corresponding to 'allene' (entry 1) and 'allenes' (entry 2) contain references where the queried term was found as a concept. Combining the two answer lists (917 hits) and then the removal of duplicates (771 citations) furnishes 630 journal articles. The expression 'N-heterocyclic carbene' (entry 8) leads to a greater hit count (668 hits) than the corresponding plural form (231 hits). A processing similar to entries 1 and 2 led to 671 citations and 508 journal articles. This emphasizes in such cases that both singular and plural forms need to be searched. With respect to the expressions 'modified nucleoside' and 'modified nucleosides' (entries 10 and 11) the largest list contains the smallest one after combining them. Because the references corresponding to the terms and expressions of entries 1, 2, 10 and 11 were selected through a concept search, the process of combining answer lists may be simplified by typing the singular and the plural forms within the same search and by using one of these forms within brackets. However this trick is not valid if the references corresponding to an expression are found containing this expression 'as entered' as in the cases of entries 8, 9, 12 and 13. Finally the term 'material' (entries 16 and 20) was searched as a concept on both databases under singular *vs.* plural forms, the resulting hit counts were found different from less than 0.05%.

The results of the expression in entry 15 are worthy of some specific explanations because we initially performed this search by selecting the expression 'copper (Cu) catalyzed arylation' found as a concept thus leading to 375 hit counts. This high value is mostly due to a high occurrence number of the term 'aryl' resulting from the truncating step of the NLQ process. In order to retrieve only chemically answers relevant to the arylation concept, we ruled out the term 'aryl' by building this query as follows: (i) references were found containing 'copper catalyzed' as entered (869 hits), (ii) references were found containing 'Cu catalyzed' as entered (254 hits), and (iii) the two answer sets were combined (104 hits). In parallel a reference list was found containing 'arylation' as entered (924 hits) and this latter hit set was intersected with the previously obtained 1044 hit set thus furnishing a final list of 76 hits. For entry 16 a similar process was set in order to get the terms 'hybrid' and 'materials' closer to each other. This query was built following the sequence: (i) references were found containing 'hybrid material' as entered (470 hits), (ii) references were found containing 'hybrid materials' as entered (780 hits), and (iii) the two answer sets were combined (1105 hits). In parallel a third reference set was found containing the concept 'nanoparticles' (39 914 hits) and this latter set was intersected with the 1105 hit count set providing 266 hits as a final result. Because all queries were performed in March, April and May 2013, the hit counts may vary slightly if performed now.

The first point we attempted to address is related to the non-negligible proportion of duplicate answers observed within the

Scifinder's answers whose total count is equal to 204 when summing all duplicates corresponding to each query. These internal duplicates were found among many Medline's articles that miss a DOI whereas the corresponding articles are assigned a DOI if the PubMed interface is queried. Among these 204 references, we observed too that some journal names are distinctly indexed between Medline and CAplus databases. Representative examples are given in Table 2. With respect to the Scopus's and WoS's databases only one and two duplicates were found respectively.

Tables 3 and 4 display the queries specific to WoS and Scopus, respectively, and the resulting hit counts related to the selected terms and expressions used within Scifinder's topic searches. Keeping in mind that Scifinder's topic searches include by default all indexing terms from titles, abstracts, index terms and supplementary terms we selected the corresponding WoS's search field TS (column 3) that covers the fields: title, abstract, author keywords and keywords Plus®. Queries to Scopus (column 3) were performed through the document search tab in basic mode together with the option gathering together title, abstract and keywords. By this way the retrieved answer lists are equivalent to the ones retrieved by using the field sum 'TITLE-ABS-KEY-AUTH' available in advanced search mode. In order to perform topic searches comparable to Scifinder's topic searches, the use of the right-hand truncation was systematically preferred because this enables a better control on WoS's and Scopus's queries. Boolean operators were also employed to target precisely all queries, especially the proximity operators, available in WoS and Scopus that retrieve the searched terms within the same bibliographic field. The WoS's operator NEAR searches terms that are distant by default at a maximum of 15 terms but this distance may be shortened. Terms within double quotes were alternatively searched as an exact expression (entries 7, 11, 12 and 17, Table 3). The logic for the proximity operator *W*/*n* is similar in Scopus. This operator requires defining a number *n* equivalent to the distance between the searched terms. The automatic truncation in Scifinder was offset within WoS's and Scopus's searches by extensive use of wildcards as exemplified in entry 1 (Tables 3 and 4) thus enabling the terms 'allene(s)' or 'allenyl' or 'allenic' to be retrieved.

2.2. Result analysis automation

All article lists (Table 1, column 7 and Tables 3 and 4, column 4) were exported as text files in a tagged format in order to analyze them and to find both common and specific references to each editor. The RIS file format was chosen as an export file format from Scifinder and Scopus while WoS's data were exported into the CIW file format. In order to quickly identify duplicates among two or three reference lists we used the Iddup script whose main instructions are described as follows. For each single input file, Iddup furnishes two text files in the RIS format, the first file contains unique articles (Table 1, column 8; Tables 3 and 4, column 5) while the second file contains duplicate references. When analyzing two different input files, Iddup identifies first internal duplicates in each list, discards them and then compares pair to pair the remaining references of the two lists. As output files, Iddup provides a file containing common references and two files containing specific references from each input file.

Table 2 Different indexed journal titles between CAPlus and Medline

Article count	CAplus	Medline
18	*Acta Crystallographica, Section E. Structure Reports Online*	*Acta crystallographica.chrom Section E, Structure reports online*
39	*Angewandte Chemie, International Edition*	*Angewandte Chemie (International ed. in English)*
86	*Chemistry -- A European Journal* *Chemistry – A European Journal*	*Chemistry (Weinheim an der Bergstrasse, Germany)*

Table 3 Queries and results from WoS

Entry	Queried expression in Scifinder	Queried expression in WoS	Articles[a]	Unique articles[b]
1	Allene(s)	TS = allene* OR TS = alleny* OR TS = alleni*	503	466
2	Organocatalysis	TS = organocataly*	997	963
3	Peptidomimetics	TS = peptidomimetic*	282	257
4	Agostic interactions	TS = (agostic NEAR interaction*)	65	63
5	Battery electrodes	TS = ((battery OR batteries) NEAR electrode*)	1068	1008
6	Graphene biosensors	TS = (graphene NEAR biosensor*)	47	47
7	N-Heterocyclic carbene(s)	TS = ''N-heterocyclic carbene*''	629	629
8	Modified nucleoside(s)	TS = (modif* NEAR nucleoside*)	117	114
9	Phosphine ligand(s)	TS = (phosphine NEAR1 ligand*)	322	319
10	Renewable feedstock	TS = (renewable NEAR feedstock*)	110	92
11	Copper (Cu) catalyzed arylation	(TS = (''copper catalyzed'') OR TS = (''Cu catalyzed'')) AND TS = arylation	105	103
12	Hybrid materials and nanoparticles	TS = (''hybrid material*'') AND TS = nanoparticle*	256	234
13	Viscosity of ionic liquids	TS = viscosity AND TS = ionic liquid*	360	350
14	Band gap in solar cells	TS = ((band NEAR gap) AND (solar NEAR cell*))	677	599
15	Statistical analyses of DNA microarrays	TS = statistical analyses of dna microarrays	93	93
16	Surface area in mesoporous materials	TS = (''surface area'') AND TS = (mesopor* material*)	207	193

[a] Article counts after filtering by year, language and document type. [b] After parsing by the Iddup script.

Table 4 Queries and results from Scopus

Entry	Queried expression in Scifinder	Queried expression in Scopus	Articles[a]	Unique articles[b]
1	Allene(s)	Allene* OR alleny* OR alleni*	359	356
2	Organocatalysis	Organocataly*	770	758
3	Peptidomimetics	Peptidomimetic*	299	294
4	Agostic interactions	Agostic *W*/15 interaction*	49	48
5	Battery electrodes	Batter* *W*/15 electrode*	1020	816
6	Graphene biosensors	graphene *W*/15 biosensor*	66	59
7	N-Heterocyclic carbene(s)	N-Heterocyclic *W*/1 carbene*	462	458
8	Modified nucleoside(s)	(modif* *W*/15 nucleoside*)	114	113
9	Phosphine ligand(s)	Phosphine *W*/1 ligand*	286	284
10	Renewable feedstock	Renewable *W*/15 feedstock*	242	154
11	Copper (Cu) catalyzed arylation	((copper *W*/1 catalyzed) OR (Cu *W*/1 catalyzed)) AND arylation	52	52
12	Hybrid materials and nanoparticles	(Hybrid *W*/1 material* and nanoparticle*)	271	235
13	Viscosity of ionic liquids	Viscosity of ionic *W*/1 liquid*	341	327
14	Band gap in solar cells	Band gap in solar cell*	606	424
15	Statistical analyses of DNA microarrays	Statistical analys* of dna microarray*	310	298
16	Surface area in mesoporous materials	Surface area in mesopor* *W*/1 material*	383	362

[a] Article counts after filtering by year, language and document type. [b] After parsing by the Iddup script.

When comparing two references from input lists without internal duplicates, Iddup assigns each pair a score that is computed based on the following filters:

- initial score = 0
- if same DOI then score = 10 (and references are identical)
- if similar title then increment score +3
- if same journal then increment score +1
- if same author count then increment score +0.5
- if similar author and same position then increment score +0.5
- if same starting page then increment score +1.5
- if same volume then increment score +0.5
- if same issue then increment score +0.5
- if scores > 5, then the two references are considered as identical.

The second instruction enables the script to overlook the next instruction in case of same DOIs are found. A similarity computing was introduced at the third instruction that compares the titles because many titles contain abbreviations or Greek characters that are not always indexed in the same way by the different editors. These statements prompted us to introduce a 12% similarity score – 12% of the length of the longest title – that was computed using the Levenshtein distance.[32] The influence of this parameter is discussed in Section 3.3. Likewise the author names present many discrepancies due to different spelling languages, typing errors or due to a different ranking in indexing their names. Our script was completed by correspondence arrays for some journal titles and for the Latin transcription of Greek characters. Finally Iddup discards citations corresponding to errata or corrections.

2.3. Comparison of reference lists

Unique articles of Tables 1, 3 and 4 are reported in Fig. 1. Overall the magnitude orders range similarly except for entries 2, 5, 7 and 14 that display higher article counts found by WoS and except for entry 15 where Scopus retrieves more articles than the other two systems. Scopus and Scifinder retrieve more articles in entries 9, 10, 15 and entries 1, 6 respectively.

These results were refined through Iddup computing by identifying the common articles (column 4 in Tables 5–7) to each pair of editors and the specific articles to each editor (columns 3 and 6 in Tables 5–7). The union of the total article counts (column 7, Tables 5–7) is given by the sum of columns 3, 4 and 6 while column 8 represents the proportion of common articles to two editors. Preliminary observations show that these proportions vary dramatically from a maximum of 80.0 to a minimum of 11.4 percent (entries 4 and 15, Table 6). Higher proportions of common articles were generally observed for single-, double- or triple-term queries than for the queries including four terms.

Though the main results were recorded in 2010, we have extended the query timespan to the years 1990, 1995, 2000, and 2005 for the four expressions: 'allenes', 'peptidomimetics', 'battery electrodes' and 'band gap in solar cells'. These expressions were selected because their corresponding queries furnished sufficient hit counts to be representative as soon as 1990. For example expressions such as 'organocatalysis' or 'N-heterocyclic carbenes' returned no answer in 1990 and 1995 and were thus discarded. A second selection criterion was based on variable lengths of these four expressions.

Full resulting data are included in the ESI† (Table S2). As general conclusions of this supplementary study, we noticed that: (i) the three databases lead to different result sets as in 2010, (ii) large non-overlapping result sets were found during the years 1990, 1995, 2000, and 2005, and (iii) the proportion of overlapping papers increases over the years except for 'peptidomimetics'.

In order to close this section, we may mention that the overall averages of shared references by Scifinder/WoS, Scifinder/Scopus and Scopus/WoS are 40.8, 46.8 and 52.2% respectively.

3. Discussion

These quite low overlaps between the three information systems may appear surprising but at least one precedent was observed in the computer sciences.[33]

3.1. Influence of term indexing

Which are the reasons why these differences are often so high? To answer this question, some reference lists corresponding to

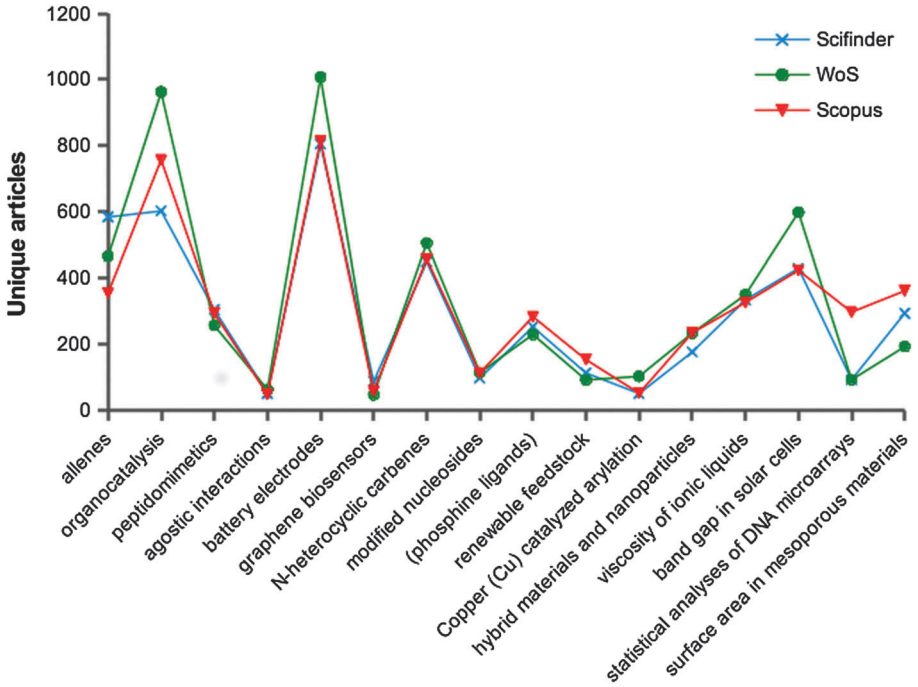

Fig. 1 Unique articles retrieved by Scifinder, WoS and Scopus.

Table 5 Iddup parsing of reference lists from Scifinder and WoS

	Scifinder			WoSz			
Entry[a]	Uniq. articles	Spec.[b]	Common[c]	Uniq. articles	Spec.[d]	Union[e]	Comm./ union[f] (%)
1	585	278	307	466	159	744	41.3
2	603	50	553	963	410	1013	54.6
3	305	123	182	257	75	380	47.9
4	51	6	45	63	18	69	65.2
5	806	328	478	1008	530	1336	35.8
6	87	46	41	47	6	93	44.1
7	450	46	404	629	225	675	59.9
8	98	30	68	114	46	144	47.2
9	253	90	163	319	156	409	39.9
10	113	47	66	92	26	139	47.5
11	51	16	35	103	68	119	29.4
12	177	50	127	234	107	284	44.7
13	335	105	230	350	120	455	50.5
14	430	248	182	599	417	847	21.5
15	93	72	21	93	72	165	12.7
16	294	232	62	193	131	425	14.6

[a] Entries 1–16 correspond to the queried expressions of previous Tables 3 and 4. [b] Specific articles to Scifinder. [c] Shared articles by both editors. [d] Specific articles to WoS. [e] Sum of columns 3, 4 and 6. [f] Proportion of common articles to two editors.

Table 6 Iddup parsing of reference lists from Scifinder and Scopus

	Scifinder			Scopus			
Entry[a]	Uniq. articles	Spec.[b]	Common[c]	Uniq. articles	Spec.[d]	Union[e]	Comm./ union[f] (%)
1	585	275	310	356	46	631	49.1
2	603	33	570	758	188	791	72.1
3	305	103	202	294	92	397	50.9
4	51	7	44	48	4	55	80.0
5	806	319	487	816	329	1135	42.9
6	87	39	48	59	11	98	49.0
7	450	49	401	458	57	507	79.1
8	98	29	69	113	44	142	48.6
9	253	75	178	284	106	359	49.6
10	113	34	79	154	75	188	42.0
11	51	14	37	52	15	66	56.1
12	177	41	136	235	99	276	49.3
13	335	100	235	327	92	427	55.0
14	430	236	194	424	230	660	29.4
15	93	53	40	298	258	351	11.4
16	294	208	86	362	276	570	15.1

[a] Entries 1–16 correspond to the queried expressions of previous Tables 3 and 4. [b] Specific articles to Scifinder. [c] Shared articles by both editors. [d] Specific articles to Scopus. [e] Sum of columns 3, 4 and 6. [f] Proportion of common articles to two editors.

specific references (columns 3 and 6) were selected and each reference of these lists was thoroughly examined in order to determine for which reason this reference was found by one editor or omitted by another one. Such reasons may be related *a priori* to journal indexing or keyword indexing but we finally found some other reasons that enabled us to assign each reference to one of the following categories:

– Journal: journal indexing may be absent or is stopped before 2010 or issue indexing is incomplete.

– Document types: Conference Proceedings, Book Reviews, and International Symposia that are not homogeneously indexed by the editors.

– Index terms: Indexing terms, Keywords and Keywords Plus®. In case of Scifinder, supplementary terms are included in index terms.

– Modified terms: (a) some journals do not provide any abstract; in those cases Scifinder designs an abstract that seems to be excerpted from the article conclusion, (b) some queried terms are

Table 7 Iddup parsing of reference lists from Scopus and WoS

	Scopus			WoS			
Entry[a]	Uniq. articles	Spec.[b]	Common[c]	Uniq. articles	Spec.[d]	Union[e]	Comm./union[f] (%)
1	356	51	305	466	161	517	59.0
2	758	44	714	963	249	1007	70.9
3	294	92	202	257	55	349	57.9
4	48	3	45	63	18	66	68.2
5	816	268	548	1008	460	1276	42.9
6	59	16	43	47	4	63	68.3
7	458	38	420	629	209	667	63.0
8	113	28	85	114	29	142	59.9
9	284	63	221	319	98	382	57.9
10	154	66	88	92	4	158	55.7
11	52	14	38	103	65	117	32.5
12	235	87	148	234	86	321	46.1
13	327	83	244	350	106	433	56.4
14	424	132	292	599	307	731	39.9
15	298	226	72	93	21	319	22.6
16	362	204	157	193	36	397	39.5

[a] Entries 1–16 correspond to the queried expressions of previous Tables 3 and 4. [b] Specific articles to Scopus. [c] Shared articles by both editors. [d] Specific articles to WoS. [e] Sum of columns 3, 4 and 6. [f] Proportion of common articles to two editors.

indexed using a hyphen included in the retrieved term *i.e.* organocatalytic, (c) the journal title is indexed in two different spellings, and (d) author keywords or titles or abstracts are modified.

– Abstracts: though provided by the journal, some abstracts are not indexed.

– Author keywords: though provided by the publisher, some author keywords are excluded from indexing.

– Different year: some issues are assigned a different year because the dates of the online publication and of the printed version are different.

– Wrong DOI: typographic errors were found in agreement with recent similar observations.[34] We noticed that a non-negligible amount of articles were missing an assigned DOI. Indeed concatenation of all articles from a particular editor followed by the removal of internal duplicates revealed that 8.7, 6.5 and 4.7% of articles from Scifinder, Scopus and WoS, respectively, were missing a DOI.

– Miscellaneous.

ESI† (Doc 1) details the whole results corresponding to the 'organocatalysis' queried term, the 'N-heterocyclic carbenes', the 'phosphine ligands' and the 'viscosity of ionic liquids' expressions. Table 8 displays the results obtained for the 'organocatalysis' queried term. The main observed differences arise from the Index Terms row. The Keywords Plus® indexing of WoS provided more articles than those retrieved by Scopus's or Scifinder's term indexing, this latter editor showing the weakest efficiency of its term indexing policy within this example. We also checked the relevance of 50 randomly selected references from the 234 references only retrieved by the Keywords Plus®. At least 45 over these 50 references were strongly related to organocatalysis. With respect to the Modified terms row, Scifinder designed an abstract excerpted from the article conclusion in one case and in the other one a hyphen was introduced in the term 'organocatalytic by WoS' (Table 8, column 3). On the same row (Table 8, entry 4, column 4), a hyphen was introduced in the term 'organocatalytic' eight times by Scifinder and in one case the term 'organocatalyst*' was shortened to 'catalyst*' within the title. Within the Abstracts row the reference found by Scifinder (Table 8, column 3) presents an abstract that was not indexed by WoS. In the case of the journal '*Angewandte Chemie, International Edition in English*', we checked 500 articles of this journal and we found that they were missing an indexed abstract by WoS. This statement is valid up to 2010 but many abstracts are indexed in more recent years. In column 4 (entry 5) the 10 references found specifically by WoS result from a left truncation of the term 'organocatalyst' to 'catalyst' in Scifinder. More surprising are the 149 references (entry 6, column 4) where Scifinder modified the original author keywords by shortening or suppressing the queried term.

Five articles were indexed by WoS with one misspelled character on their DOI compared to the original DOI (Table 8, entry 8). Finally the miscellaneous category contains articles where: (i) the filters applied to the document types during the querying step differ from one editor to another one thus during the analysis step Iddup discards citations corresponding to some unwanted document types *i.e.* book chapters and corrections, and (ii) the 0.8 similarity score on the titles and on the author names was in one case the reason why two references were wrongly differentiated.

Table 8 Study of reference lists corresponding to the 'organocatalysis' term

		Scifinder/WoS		Scifinder/Scopus		Scopus/WoS	
Entry	Category	Scifinder (50)[a]	WoS (410)[a]	Scifinder (33)[a]	Scopus (188)[a]	Scopus (44)[a]	WoS (249)[a]
1	Journals	22	0	16	4	10	5
2	Document types	5	1	5	0	4	1
3	Index terms	4	234	7	148	0	232
4	Modified terms	2	9	0	5	1	0
5	Abstracts	1	10	2	29	11	1
6	Author keywords	0	149	0	0	0	1
7	Different year	6	2	1	1	11	3
8	Wrong DOI	5	5	1	1	7	6
9	Miscellaneous	5	0	1	0	0	0
10	Checked index terms[b]	6 (7)	387 (402)	9 (9)	182 (182)	10 (12)	231 (232)

[a] Numbers within brackets correspond to specific articles reported in entries 2 of Tables 5–7. [b] Numbers within brackets correspond to the sum of articles from the indexing categories 3 to 6.

If we consider all articles of a particular editor that are classified in the index terms or modified terms or abstracts or author keywords categories, the next question remains to verify whether the concurrent editor's database is really missing this specific information or not? To check this hypothesis we injected the DOIs or the bibliographic data of a given editor's articles corresponding to the above-mentioned indexing categories into the query interface of the concurrent editor. The results are displayed in the last row (Table 8, entry 10). For example 6 over 7 specific articles retrieved by Scifinder (Table 8, column 3) are also present in the WoS thus emphasizing the importance of the Scifinder's indexing policy in this case. Once this statement has been established we noted that only 22 articles (Table 8, entry 1) from specific journals and 5 articles (Table 8, entry 2) from the document type category belong specifically to Scifinder. The vast majority of articles retrieved by WoS (Table 8, entry 10, column 4) would have been retrieved likewise by Scifinder if different indexing rules have been applied.

Comparing Scifinder and Scopus (Table 8, columns 5 and 6) on their specific references led to similar observations. The coverage of journals is in favour of Scifinder whereas Scopus retrieves a higher article count owing to its term indexing. Moreover Scopus indexes in the case of 3 reviews not only the abstracts but also the tables of contents where the queried term is present. We noticed too that author keywords were neither suppressed nor modified.

By comparing Scopus and WoS (Table 8, columns 7 and 8), we observed that WoS shows a high count of articles retrieved by the Keywords Plus® indexing. Among the 11 articles included in the abstracts category (Table 8, column 7) 3 reviews are present indexed by Scopus within their tables of contents. The 8 remaining articles of the abstracts category correspond to references for which WoS did not index the abstract. We observed that the different year category displays a rather important amount of articles: 11 articles are indexed by WoS in 2009 or 2011 and 3 articles are indexed by Scopus in 2009. Obviously these articles would have been retrieved by a multiple-year query. In the wrong DOI category were found the same articles as previously noticed.

In order to confirm the results displayed in Table 8, we analyzed some data from two-term queries and a three-term query (Table 9). The first studied expression was 'N-heterocyclic carbenes' (Table 9, columns 3 and 4) and the articles retrieved by Scifinder and Scopus respectively. Here again the influence of term indexing is predominant but to a smaller extent than previously. Within the modified terms category we observed that in some cases Scifinder developed the NHC acronym to 'N-heterocyclic carbenes' thus enabling the corresponding article to be retrieved. Finally 6 over 7 articles present in the miscellaneous category (Table 9, column 3) correspond to misspellings or typographic errors from Scopus.

The next results concerned the two-term expression 'phosphine ligands' and the retrieved articles by Scopus and WoS (Table 9, columns 5 and 6). Apart from the predominant influence of term indexing by both editors, Scopus offers in this case a slightly better journal coverage and a better abstract coverage. In the miscellaneous category Scopus retrieved some articles containing the expanded forms of the 'phosphine' term such as 'bisphosphine' or 'triphenylphosphine'. Finally we looked at the three-term query 'viscosity of ionic liquids' (Table 9, columns 7 and 8) and examined the specific articles retrieved by Scopus and WoS. The observed proportions within the different categories are similar to those obtained in previous cases, the index term category remaining the main differentiating one.

These last results (Tables 8 and 9) were not computed by any algorithmic process and only affect a part of the study presented in Tables 5–7. Nevertheless they reveal some interesting trends about the scope and the limits of term and keyword indexing policies of Scifinder, Scopus and WoS. If we focus now on the values displayed in different columns of entry 10 (Tables 8 and 9), we observe that a high proportion of articles retrieved in the indexing categories by a particular editor are present in both other editor's databases. Ultimately this emphasizes the influence of term and keyword indexing policies of these editors because most informative articles are shared by the three editors. In other words the proportion of information specific to a given editor is not as high as it could be expected from preliminary results displayed in Tables 5–7. Moreover the term and keyword indexing policies clearly

Table 9 Study of reference lists corresponding to the expressions 'N-heterocyclic carbenes', 'phosphine ligands' and 'viscosity of ionic liquids'

		N-Heterocyclic carbenes		Phosphine ligands		Viscosity of ionic liquids	
Entry	Category	Scifinder (49)[a]	Scopus (57)[a]	Scopus (63)[a]	WoS (98)[a]	Scopus (83)[a]	WoS (106)[a]
1	Journals	7	13	8	2	3	4
2	Document types	2	0	2	1	2	0
3	Index terms	15	25	31	94	72	93
4	Modified terms	9	14	1	0	0	0
5	Abstracts	3	0	13	0	2	0
6	Author keywords	0	0	0	0	0	1
7	Different year	6	3	0	1	0	1
8	Wrong DOI	0	0	0	0	2	1
9	Miscellaneous	7	2	8	0	2	6
10	Checked index terms[b]	27 (27)	39 (39)	45 (45)	91 (94)	71 (74)	88 (93)

[a] Numbers within brackets correspond to specific articles reported in entries 2 of Tables 5–7. [b] Numbers within brackets correspond to the sum of articles from the indexing categories 3 to 6.

Table 10 Influence of the citation impact on the retrievability of references

Expression		Scifinder/WoS			Scifinder/Scopus			Scopus/WoS		
'Organocatalysis'		Specific (50)[a]	Common (553)	Specific (410)	Specific (33)	Common (570)	Specific (188)	Specific (44)	Common (714)	Specific (249)
	Two best citation counts	442	707	496	75	707	243	442	707	496
		298	390	280	51	442	197	243	390	280
'N-Heterocyclic carbenes'		Specific (46)	Common (404)	Specific (225)	Specific (49)	Common (401)	Specific (57)	Specific (38)	Common (420)	Specific (209)
	Two best citation counts	200	445	1320	89	445	755	755	445	1320
		82	233	649	80	233	395	395	388	649
'Phosphine ligands'		Specific (90)	Common (163)	Specific (156)	Specific (75)	Common (178)	Specific (106)	Specific (63)	Common (221)	Specific (98)
	Two best citation counts	182	200	259	182	200	755	755	200	259
		119	101	249	119	101	92	92	101	249
'Band gap in solar cells''		Specific (248)	Common (182)	Specific (417)	Specific (236)	Common (194)	Specific (230)	Specific (132)	Common (292)	Specific (307)
	Two best citation counts	1358	394	922	538	1358	628	1358	477	922
		538	245	590	394	245	477	628	245	590

[a] Numbers within brackets correspond to specific articles reported in entries 2 of Tables 5–7.

differentiate the three studied editors in a higher proportion than their respective journal coverages do.

3.2. Influence of the citation counts

Apart from the influence of term and keyword indexing policies, we checked the influence of the citation counts on the retrievability of references. In other words how are the most cited references distributed between specific references to an editor *vs.* the common ones? To answer this question, four expressions were thoroughly studied (Table 10). Our previous results mentioned in Tables 8 and 9, especially entries 10 including the checked indexed terms, revealed that most of the references could have been retrieved if homogeneous indexing policies would have been applied between the three editors. From this postulate we selected the WoS to extract the best cited papers by querying the DOIs of each reference list. The whole references corresponding to Table 10 are given in the ESI,† Doc 2.

If we look only at the lines corresponding to the two best citation counts, we notice that high values are found either for specific or for common references. For the 'organocatalysis' expression, the common reference lists collect the best cited papers but for the three other expressions the best cited paper is found in one or two specific reference lists. We are interested in these references that are assigned such high citation counts. First the reference (Table 10, column 5 and 11, line 5) displaying a 1320 citation count corresponds to the journal '*Chemical Reviews*' from the ACS editor. Any usual abstract is given by the ACS and the WoS enables to retrieve this article owing to its Keywords Plus®. We came to the same conclusion for the article that was assigned a 259 citation count (Table 10, column 5 and 11, line 11).

For the paper that was assigned 755 citations (Table 10, columns 5 and 6, lines 5 and 11), Scopus designed its own abstract from the conclusion of the original paper – still an article from '*Chemical Reviews*'. Finally the paper, that was assigned 1358 citations (Table 10, columns 3, 7 and 9, line 15), was retrieved as a common reference by Scifinder and Scopus that both designed their own abstract containing the queried expression. These abstracts were however different. The WoS did not retrieve this paper because it seems to not design abstracts from scratch.

General trends may not be concluded from these few results, but highly cited papers are retrieved by all three editors. Moreover abstract and keyword indexing play a non-negligible role within these examples.

3.3. Influence of the similarity computing

We introduced a similarity parameter in the Iddup script – the Levenshtein distance – that may affect the duplicate counts when running either on a single reference list or a double one. This influence was studied by introducing increasing discrete values from 3 to 6, 9, 12, 18, 24 and 30% successively. In view of the great proportion of these duplicates obtained from Scifinder we concatenated the whole reference lists from this editor, the resulting file containing 5193 references. Submitting this file to Iddup computing furnished 243, 263, 269, 269, 270, 271 and 271 duplicates respectively. A maximum is reached approximately for a 24% threshold. Using a 12% threshold for the Levenshtein distance – the selected value all along this study – returned 270 duplicates and thus was different from just one unit of the maximum value (271).

Moreover a test was performed on the 'organocatalysis' expression and the results are summarized in Table 11. They correlate with the results of the previous paragraph but with an inverse trend. Low values of the threshold furnish a higher count of duplicates between Scifinder and another database because the count of unique articles from Scifinder is higher for low values of the Levenshtein distance. No difference was observed when comparing the Scopus's and the WoS's reference lists. It is worthwhile mentioning that the observed values for the common references (columns 3 and 6) as well as for

Table 11 Influence of increasing Levenshtein distances on some duplicate reference lists

Levenshtein distance (%)	Scifinder	Common	WoS	Scifinder	Common	Scopus
3	52	553	410	35	570	188
6	51	553	410	34	570	188
9	50	553	410	33	570	188
12	50	553	410	33	570	188
15	50	553	410	33	570	188
18	50	553	410	33	570	188

the WoS's and Scopus's specific references are stable. In a second time the variation of the Levenshtein distance only affects Scifinder's specific references for values less than or equal to 9.

4. Conclusions

Topic searches in chemical information systems are expected to return precise answers and we attempted to show in the first section of this paper how it can be challenging to query the web interfaces of Scifinder, Scopus and WoS using the most suitable syntax. If the personal learning involvement is shorter when starting a topic search with Scifinder, the higher precision of WoS's and Scopus's query languages may justify a slightly higher learning period. Crude results of these topic searches using simple terms or expressions up to four-term queries show rather uniform trends of the three information systems in retrieving large reference lists but with a noticeably greatest hit count retrieved by WoS over the whole answer sets. This feature results from a combined citation and semantic indexing affording new indexed terms that really expand capacities of topic searches. Though the coverage of common references retrieved by Scifinder, Scopus and WoS was shown to be incomplete owing mainly to keyword indexing (and to journal indexing though to a lesser extent), most of the references are shared by the three information systems including highly cited papers. Ideally they should be queried to get exhaustive answer lists or they should combine the powerful capacities of reliable thesauri and citation computing.[35]

Acknowledgements

Elsevier's French managers are gratefully acknowledged for having provided full access to Scopus's content during the period of March to April 2013.

References

1 D. D. Ridley, *Trend. Anal. Chem.*, 2001, **20**, 1–10.
2 K. M. Whitley, *J. Am. Soc. Inf. Sci. Technol.*, 2002, **53**, 1210–1215.
3 J. Li, F. Burnham, T. Lemley and R. M. Britton, *J. Electron. Resour. Med. Libr.*, 2010, **7**, 196–217.
4 M. E. Falagas, E. I. Pitsouni, G. A. Malietzis and G. Pappas, *FASEB J.*, 2008, **22**, 338–342.
5 E. Zass, *Heterocycles*, 2010, **82**, 63–86.
6 References – CAplus – Worldwide coverage of many scientific disciplines all in one source: http://www.cas.org/content/references.
7 References – CAplus – Core Journals: http://www.cas.org/expertise/cascontent/caplus/corejournals.html.
8 In the non-academic STN version, indexing terms are reachable by querying the CA Lexicon®, a controlled term thesaurus. Thus expanding a searched term opens the query to further related terms. For a historical point of view the interested reader could refer to: D. Zaye, W. Metanomski and A. Beach, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 392–399.
9 D. D. Ridley, *Information Retrieval: SciFinder*, John Wiley & Sons Ltd, Chichester, 2nd edn, 2009, p. 39.
10 CAS Roles in CA[SM]/Caplus[SM]: http://www.cas.org/ASSETS/EB85B919049C4E448DCF8D391788F0DD/casroles.pdf.
11 CAS REGISTRY – The gold standard for chemical substance information: http://www.cas.org/content/chemical-substances.
12 D. D. Ridley, *Chem. Aust.*, 1996, **63**, 367–369.
13 Pubmed: http://www.ncbi.nlm.nih.gov/pubmed.
14 Pubmed – Medical Subject Headings: http://www.nlm.nih.gov/mesh/filelist.html.
15 The Thomson Reuters journal selection process: http://wokinfo.com/essays/journal-selection-process/.
16 Science Citation Index Expanded: http://ip-science.thomsonreuters.com/mjl/scope/scope_scie/.
17 The KeyWords Plus® are index terms introduced by Thomson Reuters in which the terms are algorithmically derived from the titles of articles cited by the author of the article being indexed. E. Garfield and I. H. Sher, *J. Am. Soc. Inf. Sci.*, 1993, **44**, 298–299.
18 Personal communication of G. Rivalle, Strategic Business Manager at Thomson Reuters.
19 Elsevier – Scopus: http://www.elsevier.com/online-tools/scopus.
20 Elsevier – Scopus's content overview: http://www.info.sciverse.com/scopus/scopus-in-detail/facts.
21 Elsevier – Content/Database overview: http://www.elsevier.com/online-tools/engineering-village/contentdatabase-overview.
22 Elsevier – Embase: http://www.elsevier.com/online-tools/embase/about/emtree.
23 Elsevier – Geobase: http://www.elsevier.com/elsevier-products/geobase.
24 R. Nelson, W. Hensel, D. Baron and A. Beach, *J. Chem. Inf. Comput. Sci.*, 1975, **15**, 85–94.
25 S. M. Cohen, D. L. Dayton and R. Salvador, *J. Chem. Inf. Comput. Sci.*, 1976, **16**, 93–99.
26 K. Baser, S. Cohen, D. Dayton and P. Watkins, *J. Chem. Inf. Comput. Sci.*, 1978, **18**, 18–25.
27 If the notion of concept in Scifinder is not clearly defined by the CAS, the definition of a concept heading may be found within the Help index as follows: a concept heading is one of the index terms that make up an index entry in CAplus[SM]. Scifinder searches the controlled term (CT) field in CAplus for concept heading information, which characterizes the general subject matter of the reference.

28 J. Williams, *Online*, 1995, **19**, 60–66.
29 A. Ben Wagner, *J. Chem. Inf. Model.*, 2006, **46**, 767–774.
30 Except for the NOT Boolean operator that really excludes a specific term from queried terms.
31 A more comprehensive list of Scopus's databases is available within the Scopus Content Coverage Guide.
32 V. I. Levenshtein, *Sov. Phys. Dokl. (Engl. Transl.)*, 1966, **10**, 707–710.
33 A. Cavacini, *Scientometrics*, 2015, **102**, 2059–2071.
34 F. Franceschini, D. Maisano and L. Mastrogiacomo, *Scientometrics*, 2015, **102**, 2181–2186.
35 J. Qin, *J. Am. Soc. Inf. Sci.*, 2000, **51**, 166–180.