Mirror, mirror on the wall, squishy and soggy, 2 nanos tall: Strategies, methods and tools for searching homogeneous catalysts – An EPO perspective (Part 2. Non-patent literature and conclusions)☆

Matthias W.E. Goebel☆

Directorate 2104, Industrial Chemistry, The European Patent Office, D-80298 München, Germany

ABSTRACT

This paper provides an overview of various search strategies, methods and tools for searching catalysts comprising hydrides, coordination complexes or organic compounds. These are also commonly referred to as “homogeneous catalysts”, in contradistinction to true heterogeneous metal or metal oxide catalysts, typically on an inorganic support.

The characteristic material and search aspects are illustrated by way of exemplary patent applications for catalysts/systems that find use in industrially important reactions such as hydroformylation, hydrocyanation, (asymmetric) hydrogenation, olefin metathesis, cross-coupling or come from emerging fields (e.g. C–H activation, “green chemistry”). EPO internal, commercial, as well as some freeware tools are explained, including an evaluation of strengths and weaknesses.

Emphasis is placed on how these catalysts are searched at the European Patent Office, i.e. patentability searches of patent applications. Nevertheless, the information and evaluations provided herein should enable the reader to set up a strategy for any specific search problem in this field, including high precision (quick) and high recall clearance or validity searches.

Additionally, much insight is provided on search methods at the EPO, databases etc, which can be of wider interest than catalyst searches alone.

The paper is divided into two parts. This part (part 2) deals with non-patent literature searching and the overall conclusions, whereas part 1 in the previous issue of this journal dealt with terminology and the basics of patent searching. See also the Introduction in part 1 for the distribution of content between the individual parts.

2.5. Non-patent literature searching

For those NPL documents neither having been directly classified in ECLA (see Section 2.4 in part 1 of this paper) nor cited during search, examination or opposition of classified patent documents (see Section 2.4.2.1 in part 1 of this paper), several retrieval methods can be used. The choice will depend on the actual subject-matter to be searched and the desired degree of recall and precision. Three principal methods are outlined here. Since most of the databases presented here contain patent records as well, their significance for patent searching is also briefly mentioned.

2.5.1. Searching by structures and controlled terminology

This kind of searching bears resemblance to (patent) class searching, the equivalence of a class (or sub-group) being e.g. a set of defined parameters for a certain (sub)structure or a certain controlled term for a specific type of compound.

Structure searching is possible inter alia within CAS Registry or the CrossFire databases, whereas controlled terminology searching is possible with the CA files, using the built-in thesaurus “CA Lexicon”, or with the Encompass databases.

A good balance of precision and recall is typically achieved, for this reason, patent offices including the EPO, often consult one or more of these databases.

2.5.1.1. CAS

In general, the CAS files have the broadest coverage available of patent and non-patent literature, since more than 10,000 major scientific journals worldwide are abstracted. An average of
21,000 records is added per week (August 2007, still increasing) and entries are usually in the database within two weeks of publication. Efficient “one-stop” searching is therefore possible [46].

Structure searching, for ease of use typically via the STN Express interface, is essentially a novelty only search, albeit a very powerful tool for that purpose, in particular when searching broad (but sufficiently well-defined) Markush formulae. Structure searching within the CASREACT database [47] moreover enables efficient chemical reaction searching. Further details with respect to setting up structure search queries are beyond the scope of this article, this information is however available from CAS–STN [48].

Classification is given by controlled terms (/CT) and index terms (/IT). The CA Lexicon uses the controlled terms to organize the subjects in hierarchical lists, each /CT of a given subject being in a defined hierarchical relationship with another /CT, expressed by inter alia the codes NT\textsubscript{n} (narrower term of level n), BT\textsubscript{n} (broader term of level n), RT (related term) and UF (used for synonymous, but non-preferred other /CT). The subjects are extremely varied, ranging from biology via microbiology to chemistry and including materials per se, as well as uses, processes, and methods, e.g. functional definitions such as “ionic liquids”. The CAS assigned roles (/RL in general or specific three-letter codes such as /CAT for catalyst use or four-letter codes for more general aspects such as /PREP for preparation in general) are further powerful indexing terms, valuable in case of filtering large answer sets. Both answer sets (L numbers) and specific compounds can be searched in this field.

Thus, by varying the hierarchical levels and/or related terms, together with the use of keyword-like index terms, recall and precision may be influenced, in any case the retrieval of both hits and close analogues is possible, i.e. both of the patentability criteria of Novelty and Inventive Step (obviousness) can be checked, according to Articles 52(1), 54 and 56 EPC [13].

As an example, consider the following “real life” patent claim:

“Use of an ionic liquid as a solvent in a base-catalysed chemical reaction, wherein the ionic liquid is represented by the formula:

\[
\text{Cat}^+\text{X}^-
\]

wherein Cat\textsuperscript+ is selected from ammonium, phosphonium, borate, pyrazolium, DBU and DBN; and X\textsuperscript- is a sulfonate, phosphinate or halide anionic species, and characterized in that the ionic liquid is base stable.”

Fig. 10 shows an example of how controlled terminology may be applied to retrieve records for such subject-matter in the database CAPLUS, hourly rate (“HCAPLUS”, dark green highlight). This pricing option with a higher hourly fee but free search statements is advised when using the thesaurus, which generates many inter-

| L1 | (263361)SEA FILE=HCAPLUS ABB=ON PLU=ON CATALYSIS+NT, PPT, RT/CT |
| L2 | (7826)SEA FILE=HCAPLUS ABB=ON PLU=ON "IONIC LIQUIDS"+RT, PPT/CT |
| L3 | (470)SEA FILE=HCAPLUS ABB=ON PLU=ON L1 AND L2 |
| L4 | (226)SEA FILE=HCAPLUS ABB=ON PLU=ON L3 AND (IONIC LIQUIDS)/IT |
| L6 | 25 L5 (L) (BASE OR BASES OR NET3 OR TRIETHYLAMINE OR PIPERIDINE OR ?MORPHOLINE OR DMAA OR NMI OR PROLIN# OR PIPER? OR ?ETHYLPIPER? OR HYDROXIDE# OR CARBONATE# OR ACETATE# OR NAOH OR KOH OR K2CO3 OR LIOH) |
| L7 | 17 L5 (L) (LIME OR (CA 2W OH) OR KF OR LDA OR LITHIUM OR POTOASSIUM OR SODIUM OR NIF2 OR K3PO4 OR NA2CO3 OR DBU OR DBN OR DABCO) |
| L8 | 34 L6 OR L7 |
| L9 | 257 (223436-97-3 OR 223456-99-5 OR 223437-05-6 OR 223437-11-4 OR 57774-90-0) /BN |
| L10 | 55 (58875-50-6 OR 330671-30-2 OR 108259-90-1 OR 223437-10-3 OR 15312-12-6 OR 15302-90-6 OR 58875-50-6 OR 344774-04-5 OR 374683-43-9 OR 377800-52-9 OR 160745-20-2 OR 50877-49-1) /BN |
| L11 | 306 L9 OR L10 |
| L13 | 1 L12 (L) (BASE OR BASES OR NET3 OR TRIETHYLAMINE OR PIPERIDINE OR ?MORPHOLINE OR DMAA OR NMI OR PROLIN# OR PIPER? OR ?ETHYLPIPER? OR HYDROXIDE# OR CARBONATE# OR ACETATE# OR NAOH OR KOH OR K2CO3 OR LIOH) |
| L14 | 0 L12 (L) (LIME OR (CA 2W OH) OR KF OR LDA OR LITHIUM OR POTASSIUM OR SODIUM OR NIF2 OR K3PO4 OR NA2CO3 OR DBU OR DBN OR DABCO) |
| L15 | 35 L8 OR L13 |

Fig. 10. CAS: searching for certain ILs used in base-catalysed reactions.
mediate search statements from the input queries. The controlled terms “ionic liquids” and catalysis (/CT) are highlighted red, for maximum recall they are searched to include all narrower terms (NT), related terms (RT) and preferred terms (PFT). The answers are AND’ed to the claimed applications (highlighted pink) and further to individual bases mentioned (highlighted crimson), as are likewise the answers from querying known registry numbers (/RN) of ILs (highlighted light green). The two answer sets are finally OR’ed to give the overall 35 answers (cf. L15).

One drawback of CAS structure searching is that patents are only retrieved if the input (sub)structure is actually the subject of a worked example, hence close analogues are not found (i.e. a “novelty only” search is de facto performed). However, for the same reasons, the precision is therefore very high.

Another general drawback is that the nomenclature to be used in the queries deviates appreciably from that commonly accepted (IUPAC), as well as that the /CT indexing is sometimes incomplete, with terms changing over time; errors between structures and actual disclosure also occurring due to ambiguous nomenclature sometimes being used (e.g. “borate” for both true borates ([BOR4]/C0) and boranates ([BR4]/C0)). Trial and error is therefore often needed before a query runs satisfactorily with a reasonable degree of confidence.

Inter alia for these reasons, cost will be an issue for extensive searches. Thus, a structure search alone costs about EUR 150, retrieving the full reference list in e.g. CAplus as well will total at least about EUR 250.

2.5.1.2. EnCompassLIT and EnCompassPAT

These sister databases were formerly known as APILIT and API-PAT, respectively. The former covers exclusively NPL and the latter exclusively patents. They are now owned by Elsevier and have been joined with several other databases in one common web-browser search interface called Engineering Village [49]. On this interface, the title databases can be searched separately or freely combined with any other available database, by merely ticking the appropriate database boxes (shown for the combination EnCompassLIT, EnCompassPAT and Compendex in Fig. 11, highlighted yellow). The result set can subsequently be automatically de-duplicated.

The title databases are also available via the hosts STN, Dialog or Questel-Orbit (offering essentially the same search functions, but different display, saving and pricing formats). The latter host is exclusive for the related database WPAM, i.e. the Derwent WPI patent abstracts database [27], merged with APIPAT indexing.

EnCompassLIT and EnCompassPAT have a database coverage of catalysis of about 11% and 24%, respectively (2000), dealing essentially with petroleum related chemistry but also containing records relating to e.g. metallocenes, olefin oligomerisation and metathesis catalysts. They can be searched using the EnCompass Thesaurus, on the basis of which the original records are indexed. An additional valuable feature is the possibility of linking attributes to controlled terms, such as (a) particular constituent metal(s) to the controlled term “catalyst”, or the uniformity attribute “homogeneity” to the controlled term “catalysis” (see Fig. 12, red and green highlights).

On the Engineering Village “quick” or “expert” search interfaces, the electronic version of the Thesaurus can be found, by clicking the “Controlled term” link in the “Browse indexes” box (Fig. 11, red highlight). If only EnCompassPAT and/or EnCompassLIT are selected as databases, only the respective controlled terms are listed in the appearing window (Fig. 11, green highlight); including

1 For interpretation of the references to colour in legends of Figs. 11–13, the reader is referred to the web version of this article.
Compendex (as shown in Fig. 11, yellow highlight) gives also the further Compendex-specific controlled terms, for each of the controlled terms the applicable database(s) being indicated in parentheses. These may then be used directly in the query by ticking the associated checkbox (Fig. 11, blue highlight), giving a default OR-linked query, which OR operators can then be freely adapted.

**Fig. 12.** EnCompass: searching for homogeneous ruthenium catalysts.

**Fig. 13.** EnCompass: exemplary answer of a search for homogeneous ruthenium catalysts.
to the individual search problem. Different controlled terms for different databases can be used in a common search box, e.g. "catalysts" for Compendex and "catalyst" for EnCompassLIT and EnCompassPAT.

The strengths of the title databases are a deep indexing resulting in high precision, similar to CAS and including also macroscopic properties such as solubility, as well as extensive coverage since 1964, -LIT containing >860k and -PAT >450k records. Both databases are updated weekly, >25k records being added annually to EnCompassLIT and > 22k records being added annually to EnCompassPAT (May 2008). As for patents, those from BE, CA, FR, IT, NL, ZA, GB, US, DE and USSR/RU have been included from the beginning, followed by those from JP in 1968, those from EP and WO in 1979 and finally by those from the rest of the now 40 patent offices throughout the world in 1983. With respect to NPL, over 1500 scientific journals, trade magazines, conference papers and technical reports from 20 countries, in 10 languages are included as sources (300 primary, 1200 secondary sources). Since 1972 the abstracts from WPI or CAS have been used.

Especially, on the Engineering Village platform, both databases also contain efficient query and display options. Thus, queries can be refined by facets, i.e. the used databases (if more than one), authors, author affiliations, controlled terms, publication language, publication year, or any other freely selectable term. In the panel provided next to the answer set to this effect, a listing of respective field content is provided for each facet (e.g. author 1, author 2, etc.), typically sorted according to frequency of occurrence. The field contents can then be used individually or in any combination to refine the search accordingly, by limiting to (including), or excluding such content (e.g. a particular author). This panel is shown further down in Fig. 14, Section 2.5.1.3.

Irrespective of the platform, template displays can be activated in the individual patent answers to avoid pages-long index term listings (see red highlight in Fig. 13). Freely selectable search histories, automatically saved during a session, or individual answers can be saved on-site in multiple custom folders, downloaded (e.g. to a citation manager) or e-mailed for later processing. As a further convenience, direct links are provided to the esp@cenet, USPTO and Scirus (Section 2.5.2.1) interfaces.

The strength of the related WPAM database lies in the combination of extensive classification by the Derwent coding system ("Manual codes" /MC) with EnCompass indexing; furthermore the coverage is impressive, over 9 M documents being contained in 2000. In other words, both recall and precision are high. The weaknesses are that, as said, EnCompassLIT and EnCompassPAT are restricted to petroleum and fuel-related chemistry (although including e.g. hydrogen and/or CO chemistry), with changing coverage rules in the past, hence recall suffers. However, the high precision compensates partly for this drawback, since good leads can often be retrieved, for more thorough searching in other databases, e.g. esp@cenet, CAS, etc. Cost is also an issue, especially so with the host-only premium product WPAM. A further issue with the latter is the absence of non-patent literature.

The following examples, dealing with ruthenium catalysts, serve to illustrate some of the features of these databases. In each case, EnCompassLIT and EnCompassPAT were consulted in parallel. An exemplary query for ruthenium catalysts in general retrieved over 12k records. This answer set may be easily restricted to the homogeneous catalysts here at issue, as shown in Fig. 12, by simply linking the controlled term “homogeneity” (green highlight) to the original query.

![Fig. 14. Compendex: searching for ruthenium catalysts in olefin metathesis.](image-url)
With the query underlying Fig. 12, both supported heterogeneous and true homogenous catalysts are retrieved, now in a 290-member answer set (as shown: 177 from EnCompassLIT and 113 from EnCompassPAT), e.g. Ref. [54] of the answer set (not shown) relates to a normally homogenous olefin metathesis catalyst complex tethered to preferably silica (by Amir Hoveyda et al.[50]).

An exemplary answer of the search underlying Fig. 12 is shown in Fig. 13. As noted above, the catalyst constituents are presented in the form of a compact template (red highlight), with the automatically generated (“link term”) variable v1 (not shown) representing the listing of preferred metals as detailed in the abstract. All controlled terms are also shown (yellow highlight, including the here used ones). Further useful data is the Derwent accession number for patents (green highlight), which enables further searching in e.g. the WPI database (using e.g. the said Derwent Manual Code classification system), and IPC classes (blue highlight).

2.5.1.3. Compendex

This classified abstracts database, also available in the mentioned Engineering Village search interface [49], is exclusively devoted to NPL. It is also available from the hosts mentioned in the preceding section. Compared to the hosted versions (or e.g. CAS online), a fairly low cost flat rate (12/2008: $49 per 24 h) is available for searching on the Engineering Village platform.

Compendex, according to the database owner, is the most comprehensivie bibliographic database of engineering research (in its broadest sense) available today, containing over ten million references and abstracts taken from over 5600 scholarly journals, trade magazines, conference proceedings and technical reports. The broad subject areas of engineering and applied science are comprehensively represented. Online coverage is from 1969 to the present. Approximately 650,000 new records are added to the database annually from over 190 disciplines and major specialties within engineering, with weekly updates.

While many controlled terms exist in this database as well, they are not as useful for the here subject catalysts as those from EnCompassLIT/EnCompassPAT. Furthermore, linked terms and differentiation of controlled terms with respect to the role of a given material (e.g. reagent or product) do not exist in Compendex (see “Search Codes” box in Fig. 11).

Fig 14 can serve as rough comparison to the results found with EnCompassLIT. Since the controlled terms and field combination possibilities in both databases differ, no exact comparison was attempted; rather the search here was directed to any catalyst, but in the field of olefins and relating to metathesis (red highlight). Many homogenous ruthenium catalyst references are retrieved, as can also be seen in the first answers of the set as appearing in Fig. 14, the original full-text articles of which can further be inspected by clicking the yellow “full-text” button within the answer of interest. However, such links will appear only if the publisher has made the journal or article available online, and if the CrossRef service provides links to that particular publisher's holdings. The total number of NPL records retrieved is 287.

Combining the query underlying Fig. 14 with that of Fig. 12, followed by suitable facet refinement, could be useful when looking for ruthenium catalysts typically used in olefin metathesis reactions. The advantages of both the linking possibilities of EnCompassLIT/EnCompassPAT and the broader coverage of Compendex would be combined.
2.5.1.4. CrossFire

The CrossFire Database Suite comprises the databases CrossFire Beilstein [51] and CrossFire Gmelin [52]. Both databases provide comprehensive coverage, records dating back to the 1770s and are fairly inexpensive. The structures may be directly drawn for ease of use. The former is good for retrieving basic organic chemistry disclosures, i.e. corresponding to IPC and ECLA groups B01J31/00 to B01J31/10, while the latter includes organometallic compounds and coordination complexes, i.e. corresponding to IPC and ECLA groups B01J31/12 to B01J31/40.

The main drawback is the considerable delay from publication to database entry (up to several years).

2.5.2. Searching by keywords and concepts

This search method is the classic “Google-type” searching. In view of the substantial body of publications already dealing with this topic (see e.g. [3, 5, 9, 53]), the following valuable meta-search engines Scirus, Google Scholar and Google CrossRef are only briefly outlined with the aid of the resulting queries.

Despite their utility, a general problem encountered is article inaccessibility, inter alia because of non-subscription, defective or defunct links.

2.5.2.1. Scirus

The strengths of this free search engine [54] is its well-documented and very comprehensive coverage of over 480 M science-specific Web pages. Included are all science direct sources, covering hundreds of Elsevier journals, RSC and many other journals, e-prints, open access sources (e.g. PubMed), general web sources (for e.g. poster presentations), patent office sources (WIPO, USPTO, esp@cenet), digital archives (for e.g. theses, valuable for obviousness or disclosure considerations). A detailed query language is available, with result and date ranking options. The results may be refined in a facile manner by adding (consecutive) limitations offered in a list by clicking on them (see box “Refine your search” on the left-hand side in Figs. 15 and 16) or by restricting to a particular content source or file type, likewise by simple clicking (see box “Filter search results by” in Fig. 15 and, correspondingly, “Results filtered by” in Fig. 16). The selected answers can be saved, downloaded to a citation manager or e-mailed for later processing.

The few weaknesses are that the content is only very broadly classified into general science subject areas and content sources (albeit theses now being one of them), therefore only very limited browsing is possible. In other words, the search is only as good as the keywords. However, in view of the recent improvements relating to ergonomics and search refinement options, this handicap has become quite bearable.

Fig. 15 shows an exemplary search for ruthenium olefin metathesis catalysts. As can be seen, homogenous ruthenium catalysts are included, and clicking on the title opens the source document or at least the (PubMed) abstract, depending on the subscription profile of the searcher. It can also be seen that dissertations, normally only included in the “Digital Archives” section (see red highlight and Fig. 16) are not quantitatively recognized, since the blue highlighted answers in Fig. 15 are theses but do not appear in the answer set underlying Fig. 16. The content source “Other web” (green highlight) yields many poster presentations (not shown), albeit with some duplication.

Compared to the very similar and actually broader query underlying Fig. 14 (Compendex), the strength with respect to recall of Scirus is clearly seen (7783 vs. 287 answers). At the same time,
the powerful refinement options enable the precision to be raised to appreciable levels as well.

2.5.2.2. **Google scholar**

The strengths of this further free search engine [55] are its query language and advanced search features, the result ranking by relevance (albeit as determined by the search engine, not elucidated by Google), the “cited by” indication with the individual answers and the “key authors” indication at the bottom of each page, a link to each author being provided. These two types of indications are complimentary. The first type can be valuable in case of that particular answer being irretrievable or, if heavily cited, for author/inventor compilations on the specific subject-matter covered by it. The second type provides only a few authors, but these are relevant to the whole set (e.g. noble laureate R. Grubbs in the example set of Fig. 17).

As weaknesses can be mentioned the uncertain coverage, since the sources are not disclosed, the considerable “noise” from mere citations (taken from other articles), further abstracts, etc. Like Sci-rus, but even more so, the content is only very broadly classified into general science subject areas and content sources, therefore likewise browsing possibilities are very limited.

Fig. 17 again shows an exemplary search for ruthenium olefin metathesis catalysts.

Again, homogeneous ruthenium catalysts are retrieved (see answer 8). However, as can clearly be seen on this first page, the noise is considerable, answers 1, 4–7, 9 and 10 relating to citations only, answer 2 being a dead link and answer 3 being not an original article, but rather a facsimile of one stored on a university server. Although such a facsimile is clearly usable as such, it means that the original article will also be retrieved, i.e. duplication is produced. Thus only answer 8 (i.e. 10% of retrieved answers) is an original article. Evaluation of several other pages raises this percentage to an estimated value of about 20%. Considering that in total about 13k answers are to be expected and that only the first 1000 answers can be viewed, this is clearly an unsatisfactory situation, since keyword limitations might lead to loss of pertinent answers present in the original set.

Another comparative search with a smaller answer set (ruthenium catalyst articles published in nature publishing group journals) furnished similar results, namely 66 answers by Google Scholar vs. 13 answers by Scirus, i.e. again only 20% of the answers are original articles. Again the rest were found to be secondary references, such as further abstracts of actual articles or further citation(s) of actual articles in other database(s).

Of course, the further abstracts might be useful if the original article is not accessible.

2.5.2.3. **Google CrossRef**

The strengths of this yet further free search engine [56,57] are its query language and result ranking by means of the (unknown) Google algorithm. Further, due to the restriction to CrossRef participating publisher’s contents and only persistent links (DOI) being used to link to the publishers content, no noise is encountered, an advantage over e.g. Google Scholar [57].

However, this search engine, the CrossRef Search project, is still a pilot project “on hold” with no new publishers being added beyond the original 45, and no developments are planned at the moment [57]. This is in contrast to the DOI project, which has 2675 publishers currently participating [58] (12/2008), however most not for search purposes but only for cross-linking at citation level.

![Google Scholar search results](image-url)
(i.e. from the bibliography of one article to another). Notably the major publishers ACS, Elsevier and RSC are missing. Furthermore, very limited advanced search options exist, the content not being classified at all. Another sub-optimal feature is that the estimated number of answers is typically quite exaggerated and thus misleading (e.g. “about” 4570 results turn out to be just 902–903, after clicking through several screens, or, in the same fashion, “about” 4360 turn out to be 42).

In an example search, again for “ruthenium olefin metathesis” catalysts, homogeneous ruthenium catalysts are retrieved, as in the other databases above. However, the number of answers varies. Thus, the query used here (as shown in double quotes above), inputted several times, gave from 902–904 results, getting less with time. It must also be ensured that the option “repeat the search with the omitted results included” on the last page of the results is activated, otherwise only exemplary answers from the publishers are retrieved, i.e. an answer set of only around 40 in the present case (again varying, for no apparent reason, from 42 to 46, likewise getting less with time).

In any case, less answers are found than with Google Scholar, which is not surprising concerning the noise omission here. Likewise, less answers are found than with Scirus (1258 results there, limited to journal contents only, as can be seen in the content source box in the upper left-hand corner of Fig. 16), which is also not surprising since the publisher coverage of Scirus is greater. However, Scirus also contains some non-working or defunct links, while Google CrossRef does not.

Inspection of the answer set of said example search reveals that the important seminal Grubbs, Hermann and Nolan papers of 1998–1999 are only partly among them, likewise pertinent reviews on olefin metathesis with such complexes. An overview of these papers and reviews is found in “spotlight” [59]. Thus, only the articles of Refs. [3b,3c (reviews on olefin metathesis), 7b (seminal paper on mixed N-heterocyclic carbene/phosphine ligands by Herrmann et al.)] are retrieved, while those of Refs. [3a,5 (Nolan et al.), 6a (Grubbs et al.), 6b,7a,7c,8 (exhaustive mechanistic studies by Grubbs et al.)] are not (due to non-participating publishers ACS and Elsevier).

A possible work-around is outlined as follows, with the aid of Fig. 18. As shown with the blue highlight, the publication year of suspected articles may be added as a general limitation. This is recognized by the search engine as a “special” input and a series of options for a Google scholar search at the top of the first page is now displayed, namely to search the same (keyword) query in Google scholar (red highlight), or to go directly to the three most cited papers (by first author appearance) that would be found in that Google scholar search (green highlights). Once the article of interest is selected, it appears either as [citation] only, or as link to the desired article, depending on the publisher being included in the CrossRef search pilot or not (since the scholar engine is still limited to the CrossRef publishers). In the former case, the link below that answer “all ... versions” needs to be selected. Once that is done, links will appear in a new scholar window, as well as previous [citation] version. From these, the link referring to the publisher’s website in the author name line needs to be selected to give the desired original article (provided a valid subscription exists). In some cases, a link to a database might have to be chosen first (e.g. PubMed).

It is likely that, in view of the recent enhancements to Scirus (see Section 2.5.2.1), Google CrossRef will become less important as search tool, at least as long as the coverage remains at 45 publishers.

Fig. 18. Google CrossRef: crossing over to Google scholar with number restrictions.
2.5.3. Searching in specialty databases

Of help in sufficiently narrow fields of search can be databases covering just such fields; roughly speaking the whole database would thus correspond to one or only a few classification subgroups and/or index terms. A typical advantage, due to smaller size, is that (at least limited) browsing is possible. This can be an advantage if no good keywords exist for the subject-matter at issue.

For example, the database “IL Thermo” relates exclusively to ionic liquids known to date, supplying e.g. their registry numbers (for searches such as in Section 2.5.1.1) [60]. Molecular formulae, molecular weight, IL name (parts) such as the generic cation or anion can be inputted, the answer sets can then be browsed. Another example is the “Fischer-Tropsch Archive”, being a large collection of NPL and patent documents from the 1920s through the present day relating to synthetic fuel, including catalysts therefor. It is a unique site for such documents collected from various historical sources, stored in this centralized location and being available as electronic media [61].

Also of use can be (usually free) publication lists of leading authors in the particular field of interest. Again, the lists can be browsed for relevant articles, these in turn data-mined for keywords, other relevant authors, references, etc. For example, the publication list of Prof. Omar M. Yaghi, a leading author in the field of MOFs, is up-to-date and contains over 100 entries and allows browsing of the abstracts and bibliographic data, direct links being provided to open the original articles [62].

Mention may finally be made of databases especially for academic publications such as the “Digital Collections”, searchable with the Bielefeld Academic Search Engine (BASE) [63]. Currently over 15 M documents are contained from over 1000 academic content providers, mainly universities. Accessibility of full-text content is thus less of a problem than with commercial search engines. Furthermore, “deep web” publications which are ignored by commercial search engines or get lost in the vast quantity of hits can also be retrieved. Limited browsing within mainly dissertations (about 50k) is also possible within the BASE Lab subsection, accessible via the help menu.

The initial results of a keyword search are refinable by e.g. authors, resources, document type, and language. Since the document type “thesis” is recognized, this search engine can selectively retrieve such publications. Another neat feature is the implemented sequential broadening or narrowing of input search terms. Thus, plural, genitive and other word forms are searched automatically if the initial settings are used, i.e. the search box “find additional word forms” has a tick. If this is not desired, it may simply be unticked. On the other hand, if broadening is desired, the Eurovoc thesaurus may be activated further to said default tick, either to the option “basic terms only”, or to the still broader option “basic terms and used-for-terms”. With the first option, a search term will be searched in up to 21 languages at the same time (no matter which language you use), provided that the search term is included in Eurovoc (for every language 6500 basic terms are included). With the second option, synonyms for this search term will be searched additionally. All in all 239 k terms are included in Eurovoc.

An example search, again for ruthenium metathesis catalysts is shown in Fig. 19. As can be seen in comparison to the same query submitted to Google CrossRef (see Section 2.5.2.3), the number of hits (128 vs. 902–903) is lower in this academic database, as expected, further some duplication is observed (see answers 2/3). However, the three unique answers shown are all theses, as said, valuable for issues of e.g. obviousness (if not novelty) and disclosure or enablement. Although the document type only indicates “thesis” for one of these (answer 1, red highlight), answers 3 and 4 are clearly also theses (see green and blue highlights, respectively), even though the document type for 3 is indicated as “unknown” (see refinement pull-down menu “document type” on the right) and that for 4 as “text” (loc. cit.).

On the other hand, deactivating the Eurovoc thesaurus gives only 109 instead of 128 answers, e.g. the German language answer 1 being absent. This thesaurus is therefore a powerful tool.

As can furthermore be seen when comparing Figs. 15 and 19, answer 5 in the former corresponds to answer 2/3 in the latter. It is noted that Scirus did not recognize it as a dissertation either.
2.6. Problems during search: “Complex Applications”

The EPO tries to issue a complete search report for every single file it receives. However, this is not always possible. Increasingly, and for various reasons, applications are received with:

- broad (but clear) claims, e.g. A–B–C type Markush structures, which as such exceed the (quite generous) CAS Registry search limits (both as elemental formulas and as chemical structures) and are usually not supported by the description or fully disclosed therein, i.e. they do not comply with the requirements of Article 83 and 84 EPC;
- unclear claims, either by way of a “smoke screen” of independent claims, as such 30, with overlapping scope (not uncommon in US applications, due to different legal provisions and claim drafting practise there), or by way of multitudinous dependent claims, e.g. 153, or the individual claims containing unclear features, either per se or by too many permutations, options and/or provisos, or the subject-matter to be protected being defined by a result to be achieved, or by obscure parameters, these typically serving to disguise a lack of novelty, i.e. they do not comply with the requirements of Article 84 EPC;
- claims directed to well-known subject-matter, in chemistry known as “too simple compounds” like generic heterocycles, generic bridged multinuclear complexes, etc., such compounds or compositions leading to innumerable novelty destroying documents, a situation referred to as “novelty overflow”, i.e. they do not comply with the requirements of Article 84 EPC, since it is not clear which restriction is needed and/or desired to overcome the severe lack of Novelty.

Thus all of the knowledge gained above would be useless, or at least pointless, in such a situation (see also [64]). In any case the needed search effort is tremendous. A thorough discussion of “complex applications” and the deficiencies mentioned above is found in [22].

2.7. EPO’s solutions

So what does the search examiner at the EPO do? What can he do? Well, he can limit the extent of the search to that subject-matter, which is clear, supported, disclosed and/or (more likely) patentable and communicate this to the applicant. In other words, the restriction will be to the core of the invention (worked examples, description and/or dependent claims), and a “meaningful” search is the result. This is referred to as the “Complex Applications Approach”, a further discussion hub in [22].

The basis for issuing an incomplete search is Rule 63 (formerly 45) EPC [13], which basically states that when an application does not meet certain provisions of the EPC, to such an extent that a “meaningful search” is not possible, then a declaration of no search or an incomplete search may be issued. The said declaration is, of course, restricted to extreme cases, where no clear core is discernible and/or the core is obscure (e.g. a claim to Kryptonite).

Another possibility, if clearly more than one invention is involved (per se or in view of the prior art) is to use the administrative provision under Article 82 and Rules 44, 64 EPC and raise a lack of unity of invention objection, inviting the applicant to pay further search fees for each further found invention, should he wish to have these searched. A thorough discussion on this topic is found in [65].

Yet further, the two possibilities may be combined, i.e. the search is both limited and initially directed to certain section(s) of claimed subject-matter only, according to the found separate invention first mentioned in the claims.

Similar strategies exist for PCT applications, many of which are also searched by the EPO in its capacity as International Searching Authority (ISA), see Article 17(2)(a)(ii) and (b) PCT with respect to incomplete searches and declarations of no search, Articles 3(4)(iii) and 17(3)(a) PCT, in combination with Rule 13 PCT, with respect to unity of invention [66,67]; see also the PCT International Search and Preliminary Examination Guidelines, chapters 9 (incomplete or no searches) and 10 (unity of invention) [68].

The reason for drawing attention to these search-related problems in this paper is that it can be assumed that, at least in the case of some of such submitted faulty applications, such grave deficiencies might possibly result from incomplete knowledge of the prior art and/or the way prior art is used by a patent office, e.g. the EPO. Put in different terms, a “sloppy” search, together with inadequate application drafting, will result in serious problems, both for the applicant and for the patent office concerned, e.g. the EPO. Reducing the number of such applications would be a “win/win/win” situation for applicants, patent offices and third parties needing to know their freedom to operate, i.e. what their competitors claim and can be realistically expected to get.

3. Conclusions

In conclusion, the strengths and weaknesses of the various search methods and tools have been elucidated, based on example queries. From this it has become clearer why broad, unclear or clearly non-novel claims pose such a big problem to the patent community and the patent examiner at large.

It is now furthermore clear that no strategy is universal, since, as the saying goes, a jack of all trades (will be) a master of none. Thus, NPL and patent literature searches often require different, complementary strategies and the search of overly broad patent application claims may need to be restricted (the only “search tool” left if nothing else helps).

Summing up, the following general conclusions regarding suitable (combinations of) search strategies and tools may be drawn:

- well-defined homogenous catalysts:
  - structures can be searched in CAS Registry and/or CASREACT,
  - and/or ECLA classes in some cases, e.g. with:
    - N-heterocyclic carbene ligands (B01J31/22C),
    - heterocyclic N ligands (B01J31/18B and lower),
    - PXn ligands (X = O, N; n = 1–3; B01J31/18C and lower),
    - cyclic phosphines (B01J31/24 and lower),
    - mixed P/N and P/P ligands (B01J31/18E, B01J31/24D);
- generic homogenous catalysts:
  - if good keywords for the catalyst application (e.g. hydrogenation, olefin metathesis) and/or any structural aspects (e.g. ruthenium) are known, meta-search engines can be useful;
  - ECLA classes and search in title, abstract keywords, and/or
  - UCLA classes and search in full-text keywords, and/or
  - F-terms, and/or
  - CAS Thesaurus queries;
- supported (“heterogenized”) homogenous catalysts:
  - structure searching in CAS Registry, and/or
  - using the CAS Thesaurus, and/or
  - ECLA classes for coordination complexes (B01J31/16C and lower, B01J31/16D and lower, B01J31/16F), and ILs (B01J31/02G4 and lower);
  - optionally together with meta-search engine(s).
In any case, the ECLA function classes should also be checked for completeness, e.g. C07C45/50 for the preparation of aldehydes or ketones via hydroformylation, C07F15/00 for organic Group 8–10 metal complexes per se or C07F17/00 for metallocenes per se, since catalysts are not classified in some cases, e.g. old documents, documents mentioning catalysts or catalytic use only generally, or documents where exclusively a process with a generic or fairly well-known catalyst is concerned, the invention thus residing more in a modification of a process using it to make a product. In such cases, it is likely that only the product and/or its preparation process would be classified. The CAS answer sets can be narrowed with compound roles (JRL), but caution should be exercised, since this assignment may not be complete.

In many cases, additional reference hunting will be needed, both for patent and NPL cases, i.e. the references of a lead document are analyzed to find better hits. This may need to be performed several times in an iterative fashion, e.g. for retrieving old hit documents for suspected well-known subject-matter, e.g. presumably “standard” preparation procedures. Hyperlinks in the HTML versions of the online articles on the publisher's website are well-suited for this purpose. In this way, a hit document from 1934 was retrieved in an own search case after 3 iterations, starting from the “Houben-Weyl” organic chemistry collection, which had not been retrieved in the CAS or other files.

Acknowledgements

This article is based in part on the presentations prepared for the EPO seminars for applicants “Search matters 2008” and “Search matters 2009”. I am very grateful to Sytske de Jonge and Dieter Hofmann at the EPO for helpful discussions. I am also grateful to Natalzia Lewanzik and Núria Roqué-Rosell for their efforts in reclassifying the documents of B01J31 at the EPO.

References

Some of the references listed below are taken from part 1 of this article and have been included here for the purposes of completeness and unique reference numbering.


Matthias Goebel studied Chemistry at the Technische Universität München and graduated in the field of carbohydrate chemistry in 1994. After a post-doctoral stay with Peter Vollhardt at the University of California in Berkeley from 1995 to 1996 in the field of organometallic chemistry, he joined the EPO in 1997 as a substantive examiner in the chemical engineering field, now within the cluster Industrial Chemistry. Since 2000, he has been a search examiner as well, specializing in catalysts and solid sorbents. He was nominated the caretaker (“Gerant”) of the catalyst class B01J31 in 2005 and has since then been busy with reclassifying the documents therein. Among other duties, he has been the directorate specialist for complex applications since 2003 and for non-unity (protest review board) since 2006.