Trends in Professional and Academic Online Information Services

Péter JACSÓ
University of Hawaii - Department of Information and Computer Sciences, USA
jacso@hawaii.edu

INFORUM 2007: 13th Conference on Professional Information Resources
Prague, May 22-24, 2007

Since the 1970s there has been fierce competition among the major players within the traditional, subscription-based professional and academic online information services arena At the beginning, there were SDC, Dialog, BRS and Mead Data Central (Lexis-Nexis).

A few years later, several other companies have joined the quite lucrative market of making primarily large indexing and abstracting databases and various business directories online searchable. These included H.W.Wilson, InfoTrac (now part of the GaleGroup), Ebsco, Ovid (a partial successor to BRS, Bibliographic Retrieval Services), CSA, UMI (now ProQuest, and since very recently ProQuest/CSA). They were competing both in the content and the software features territory. Except ofr a watered down version

End-users who wanted to use academic and research databases (mostly indexing/abstracting ones) but were not affiliated with a university or large public library had to put up with the relatively inexpensive and absolutely “cheap”, watered down versions of BRS and Dialog, called BRS/AfterDark and Dialog Knowledge Index which were no competitors to the much more sophisticated and expensive academic and professional information retrieval services.

New era, new layers, new players

That era is definitely over, and not only students and academics but even many professional searchers are happy to go to Yahoo, Google Scholar, Ask, Live Academic, Scirus, LookSmart, HighBeam to find information for free about scholarly and professional articles and conference papers, as well as the full documents. Publishers of academic and professional journals have licensed their content (selectively and with restrictions) to aggregators, but in the past few years most of them have gone digital on their own (or though digital facilitators like HighWire Press, MetaPress, Atypon, Ingenta), and offer open access to tens of millions of bibliographic records and abstracts, and a couple of million full text articles. As of mid-May, 2007 HighWire Press alone had 1.7 million full text articles and about 3.5 million abstracts from top ranking scholarly journals for free, and a superb software. It represents the top of the competition, but subscription-based online information services will have to compete with many other powerful alternatives.

As I told a year ago in my closing address at the Annual Conference of the UK Serials Group < http://www.uksg.org/sites/uksg.org/files/imported/presentations8/jacso.ppt >, small, traditional indexing/abstracting databases who don’t innovate radically, will go extinct. I would add now in this Inforum 2007 keynote speech, where representatives of many of the largest academic and professional online services are present (along with information professionals), that even the largest aggregators with many full text mega-databases will have to keep innovating and adapting to the prevailing culture of information retrieval just to keep their subscribers base and especially if they want to increase it.

They have to bring the most and the best out of the potential synergy of having several dozen, and in some cases, several hundred databases. The only way to do it is to accept and even embrace the new style of searching, to use and improve many of the new (or at least new looking) features so prevalent in the open access search domain. Academic and professional online services still have the advantage of knowing very well the hierarchy and structure, the fine points and the idiosyncrasies of abstracting/and indexing and full text databases which are the weakest points of the software focused Windows Live Academic and Google Scholar databases. They have barely used the rich metadata which was made available to them along with the full text of 15-20 million scholarly articles by the hundreds of publishers who wanted to be eagerly incorporated in Windows Live Academic and especially Google Scholar. Google Scholar is particularly notorious of badly parsing the source text, and to take, for example, many –if not any- 4 character string to be a publication year even if it is a page number, part of a street address or the phone number of the publisher.

Lack of decent parsing of text creates in Google Scholar so many identically named and highly productive authors as I. Introduction, I. Preface or their younger siblings like II Background – at the expense of the real authors whose name as the author search element would not bring up their articles or their full text documents. As for Windows Live Academic, it is a sorry effort from the long-time leader of the commercial software revolution even for the lower spheres of academia. Both companies handled the millions of well-tagged digital documents, their bibliographic records and/or their abstracts as they had to handle the mostly unstructured, untagged gazillions of Web pages created by hundred thousands of John and Jane Does who never bothered to assign metadata to their digital creations.

That must be the reason that Live Academic does not offer at all field specific access points, such as author, publication year, journal name and the like. Microsoft removed the appallingly dirty list of journals covered by Live Academic and disabled the search for stopwords which I used to prove that at the launch they had 4 million records not six million but this does not change the impression that it is a very undergraduate academic product. Google Scholar offers somewhat more access points or so it seems from the advanced search page, but many of them should be used with great caution, because the field-specific indexes created in Google Scholar turn out to be as scholarly as high school students in their rented tuxedos at the graduation party. Google Scholar apparently also missed the class on Boolean logic but claims to use it. In it non-conformist interpretation of some basic tenets in searching (and common sense), it can’t handle date range searching even at elementary level, but offers that option and brings up absurd results. Less visibly, but very importantly, it often misidentifies articles citing the one at hand just because parts of the bibliographic elements between the two do match.

Google Scholar worshippers (bloggers and professors alike) don’t seem to recognize the oddity of being informed in mid-2007 about thousands of papers to be published not only in 2008 but also in the next decade. Yes, I am aware of early bird papers which appear digitally weeks and sometimes months before the print edition. It may be also chalked up to this worshipping mentality for the Dear Leader, but I don’t know of others who would have brought up the issue of the many phantom citations, which are obvious at least in those cases when the papers to be published in 2008 and thereafter are reported by Google Scholar to have been cited by papers published a decade ago. The number of records with absurd citedness counts must be orders of magnitude larger, but they are not nearly as obvious as these examples.

The very same worshippers unfortunately blog a lot, and even publish apparently unrefereed papers in traditional publications which look like hagiographies written by diligent public school students in neat school uniforms in countries where there is still personality cult. It is very discouraging when a professor in fisheries science writes an article about Google Scholar, http://www.int-res.com/articles/esep/2005/E65.pdf > based on some test queries and based on the hit counts and citedness scores deliveed by Google Schoalr concludes that it is equivalent to the results provided by Web of Science. If I were to dabble in fisheries science with my mighty ignorance of the mating habits of fishes in the Andaman Sea based on observations of the mating behavior of a few dozen fishes, I could not and should not publish a paper in a decent journal, but the Journal of Ethics in Science and Environmental Politics was happy to publish this “research”. I doubt that it went through any refereeing process by an expert., and would be interested in seing a paper about the ethics of publishing such a paper. Other academics also published papers without recognizing that the hit counts and citedness counts of Google Scholar are absurd and useless most of the time. They are like snake oil versus real medicine. I illustrate some of the underlying problems of the phantom citations and highly inflated citedness and hit counts dispensed by Google Scholar which can't even handle basic Boolean operations, date range searching, let alone accurate citation matching < http://projects.ics.hawaii.edu/~jacso/PDFs/jacso-deflated-inflated.pdf >.

Then there are the journalists, who are always ready to publish something about Google. The incompetent and sloppy piece done by a reporter of Nature more than a year ago about Google Scholar < http://www.nature.com/nature/journal/v438/n7068/full/438554a.html >, added an extra thick layer of undeserved credit to Google Scholar. Unfortunately, the clout earned by Nature through papers written by real scholars extends the halo effect to some of the roving reporters of Nature who deal with avian flu one day, global warming the other day, and Google Scholar the next day has the obvious consequences.   I hope that the professional and scholarly online information services will not follow the unscholarly practices of Google Scholar.

With all that said, subscription-based academic and professional online information services can’t ignore their largest competitors, because they may brain up simply by hiring an experienced developer from one of the traditional information services, and clean up the databases by re-harvesting the publishers’ digital archives (not a big deal these days) applying a rational parsing algorithm and using metadata.

Since the late 1990s some of the mushrooming and (at least partially) free Internet-based academic and professional online databases such as PubMed, Agricola, NTIS, ERIC, and TRIS posed a new challenge to the traditional players by delivering content for free. At first, these could be dismissed because of the modest software features which lacked essential options such as browsing, sorting, and proximity operators. However, some of the freebies had remarkable features that were not offered and are still not matched by most of the professional aggregators, such as query clarification and disambiguation.

In addition, most of the digital facilitators, many of the publishers, and entrepreneurs do have excellent or much improving and very large digital collections and good or excellent search engines. The digital facilitators can and do act as aggregators. Within a few weeks more than a dozen scholarly societies/associations will launch the Scitopia Web site with information about 5 million scholarly articles. Unusually, the PR material mentions only 3 million items, but I think it is a gross underestimate, just the opposite of what the Windows Live Academic developers did. Scitopia will use the software which has already proven itself with Science.gov, an excellent aggregate systems of scholarly publications of government agencies.

Dual mission: consumer and professional oriented services

The features (to be) adapted by academic and professional online information services range from better understanding and disambiguating the terms the users entered, to showing them the possibly relevant digital resources by displaying a score board of hit counts from several databases, to showing clusters of the results in order to increase the precision of the result set and to refine it, to lead the users to related materials using cited and citing reference links, and to enhance the primary data with information gleaned from other subscription-based and open access digital resources mashed up in Web 2.0 style for efficient cherry-picking by the users.

To please the traditional academics and the young, upwardly mobile Millennium Generation, academic and professional online services must adapt to the prevailing attitudes and expectations, and must adopt the best practices of the free or very inexpensive online services along with those of the time honored advanced searching options of the professional information services. Otherwise they would not be able to make profit for much longer. Some aggregators, which provide only the software platform but have no data of their own, would not even survive without adaptation as Darwin would point out. In this keynote talk I can only touch on some of the issues that I consider the most important, such as the assistance in formulating the queries, the selection of databases, providing parameters and demographic indicators about the digital collections, and also the result list to allow the users to have a feel about the territory and the lay of land, clues and signs for navigating within and among the database efficiently through following cited and citing references, see additional hints for cherry picking the most pertinent items from the result list, and discovering source documents not available in the database searched but available through other resources licensed by the library.

Minding your search words

For many of the users the problem is not merely choosing the right word for a query, but spelling it right. Still, most of the academic and professional online services don’t pay enough attention to this issue. Starbuck would not be able to make its huge profit if their barristas would not care to understand in 86 accents the multiple syllable fancy names of coffees no matter how much its clientele strains in trying to remember the right order of the adjectives for the order of the Iced Half-Caf Triple Grande Caramel Macchiato Non-fat with Whip Latte, or torture the Italian language a la Michael Angel for Michelangelo when asking for a café doppio restritto con panna supremo and invariably put the emphasis on the wrong syllables.

One of the reasons for the popularity of Google from the get go was its coyly "did you mean" question followed by the correct spelling of the word. It triggers the same "oh-how-sweet" reaction as the guy who shows up with some nice candy in the office on Valentine Day and tells that he was late because he helped a group of blind ladies across the street.

It is a huge difference from the cold "no record found" reply for the query Time Square in the New York Times full text database on Dialog where you are supposed to use Times()Square to pay homage to the oddity of the implication of using a space between query words which would be interpreted by Dialog as exact descriptor. Unless you were born genetically endowed with that idea and a silver spoon in your mouth, or had been using and teaching Dialog for decades, you would never guess it, and would feel unhappy seeing zero result for such an obviously correct query.

In the professional online systems it is exceptional when a software would gingerly help you with the correct spelling of your search word. Ebsco tries it, but sometimes it makes strange offers that would find no matches in the database, such as brow sable for the query term browsable as in browsable and seachable indexes. It is also good with automatically searching, i.e. without "did you meaning" the users, for regular plurals, British and American variants such as favor or favour, but it is not as good as Lexis, the exception to the rule, because Ebsco does not do it consistently as it happens with the word encyclopaedia which does not find the records where the word is spelled as encyclopedia, quite an unfortunate oversight in an American database.

Strangely, a government database seems to offer the best solution for misspelled words. I am impressed by the native software (and content) of the National Criminal Justice Reference Services (NCJRS) database of the National Institute of Justice. Its free version offers such a smart natural language software that is capable of recognizing even the most brutally misspelled words, such as metafetamin, or **metafetamine**, and retrieves more than 500 records which have the word correctly spelled as methamphetamine in the title, descriptor or abstract. The commercial versions of this database do not find any matching records but the native version of NCJRS does find 180 hits when searching for metafetamin in the title alone. Using the widely available lists of most commonly misspelled words any online service should give it a try to better handle pervasive misspellings and please their customers.

The choice of synonyms can have much influence on the result set, and they present even more difficulty for users. A word that works well in one database may not produce any hits in another one, or produce only very few ones. In an American database interlending would yield few results, and in a British or Canadian database interlibrary loan is not nearly as useful as interlending. Users cannot be expected to have these synonyms on the tip of their tongue, neither to go and check the thesaurus.

Let the software bring to them the power of the many good thesauri which have the non-preferred terms that the preferred term is used for. When they look for books or articles about wife abuse let the software ask gently (and with checkboxes) if they would be interested also in items about abused wives (or husbands), spouse abuse, spousal abuse (which solves the gender issue), and partner abuse – which accommodates the marital aspects.

In Australian databases chances are much better to find records for articles about paediatric anaesthesia or pupils behaviour then in an American database where pediatric anesthesia and students behavior offer better chances. Using multiple word queries (which is quite normal) can have bewildering different interpretations across systems available for users. I mentioned one earlier, but spaces may trigger exact phrase search in CSA and Ebsco, but AND searches in most other systems, and Boolen OR search in a few.

How a query of three or more words is interpreted can make a great difference. I know only ProQuest which tries to apply a common-sense attitude when the user enters multiple word queries. If it is only a two word query , such as information systems then it is interpreted as a phrase, but in case of three or more words, such as online information systems, ProQuest launches a Boolean AND search.

Searching for names of authors, institutes or journal names is far more challenging because of the endless variations in abbreviations and punctuations and because of the many errors. The only perfect solution that I know is from the Getty Foundation which does an awesome job of recognizing many variants of Flemish artists' name and transforms them into Getty's preferred variant behind the scene. It would be a large job for the mega-databases to use such a solution, but much of it can be automated. On a smaller scale, look at how good are the best airline fare aggregators in helping you to clarify whether you mean Paris, Texas, or Paris, France and if you mean the Charles de Gaulle airport for Paris, or Orly, or either. It is not realistic to expect the users to keep browsing the author names and journal title indexes to find the dozens of abbreviated variations of the same journal's name or the same author's name – not to mention the omnipresent misspellings.

Database selection

In the current professional and academic systems users must first choose a database and enter the search after the selection. I think that the whole search process should start with just accepting a rather free-form query, trying to understand what the user may have wanted to ask (as described above), and pass the query to the most likely appropriate 8-10 databases focused on the sciences, social sciences, technology, or arts and humanities (which takes les than a minute in my experience).

This approach is like a straw poll used to get an impression. The system would summarize the results of the query in a scoreboard, presenting the number of hits from the query along with the names of the 6-8 most likely useful databases for the topic as represented by the query. Most of the professional and academic online information services don't offer such straw poll searches, or limit the number of databases that a query can be submitted to. Dialog has the most powerful option (DialIndex) for sweeping searches, and compact result display, but it requires prior education and knowledge. CSA shows a good solution as typing in the search term depression and choosing the title index through a pull-down menu, it broadcasts the search and returns a scoreboard in less than a minute from 50 databases. The query can be submitted to four predetermined database clusters. There should be an option to limit the search to full text databases now that CSA already implemented the full text version of several databases after the merger with ProQuest. I wish it would offer a sort option to bring the most productive databases to the top.

The Central Search metasearch and clustering engine (now called 360 Search), provides not only a powerful metasearch engine but combines it with a licensed version of the Vivisimo clustering engine. It greatly helps the resource discovery process as well as the query refinement steps by offering clusters for major topics, authors, journals, etc. It is appealing that cluster enhanced version does not cost extra for existing users. The excellent Polymeta search engine offers similar functionality as 360 Search, and is also enhanced by an intelligent "did you mean feature". It is a product of Hungarian developers, but that is not the reason of my finding it to be a trend-setter. Other federated search engines also look at enhancements through clustering alternatives..

Giving a feel about the collection and the results

Getting some perspectives about the collection in a catalog, or articles in a huge journal archive is very important and can be very useful if it is presented with skills. I have seen some very encouraging examples for giving a feel about the lay of the land of a result set of several hundred hits (or for that matter of the mega catalog of the library) in the Springer archive which sports a new interface and search engine and in the OPAC of the North Carolina State University Library. As you progress by adding a filter, such as publication year range, the demography data will also change. I wish these filters were available with a check box, to choose more than one option, and to quickly undo the filter effect.

Enhancements of records with cited references

One of the most important improvement trends is the widening appreciation and implementation of a more than 50 years old idea of Eugene Garfield. Search by cited references much alleviates many of the problems related to vocabulary, terminology, spelling and abbreviation differences mentioned earlier. There are only a handful of databases where cited references are not merely included but tagged and indexed at the micro level (cited author, cited year, cited source title, etc.). There are huge differences in the volume of cited references and very significant differences also in the correct recognition of the matches between cited and citing items. The Web of Science system now has nearly 40 million bibliographic records. This is a pretty large database by anyone's measure, but it is enlarged by more than an order of magnitude by the 750-650-700 million cited references added to the records of approximately 32 million papers which have cited references out of of nearly 40 million total records, creating an impressive network of records interconnected through the cited references. The only other comparable citation enhanced database, Scopus has about 30 million records of which nearly 12 million are enhanced with cited references, providing a total of close to 250 million cited references. In the coming weeks 7 million records will be added to the database but these will not be enhanced by cited references. The proportions are so different because Scopus has records enhanced with cited references only from 1995 (and they modestly claim from 1996) onward.

Ebsco has started enhancing some of its databases with cited references. The largest one, Academic Search Premier has close to 13 million records. A little more than 1 million records were enhanced with cited references. The CSA Technology database has 8.5 million records, and close to half million records are enhanced by cited references. PsycINFO with 2.4 milion records as of mid-May, 2007 is a leader in the discipline-specific arena with the citation enhancement developments. It has enhanced about 600,000 records with cited references, and reports to have about 23 million cited references in PsycINFO, mostly from 1999 onward (the PR materials refers to its as comprehensive enhancement started from 2001). The 40:1 ratio between cited references and enhanced records is understandable when you realize that it has enhanced records for books which often have several hundred and some have several thousand references, as opposed to the average article with about 20 cited references. There are a huge difference how PsycINFO is implemented on the Ovid, OCLC, Dialog, DataStar and other platforms. Two hosts stand out with the most efficient implementation, CSA and Ebsco, and the former is way ahead of the latter simply because it lists the citedness score of the cited references, providing immense help in picking the most cited, and therefore most promising papers from the list of cited references. Hopefully, other prominent databases, like INSPEC will follow this trend.

Searching for scholarly articles with illustrations

Web search engines have been offering image searches for many years. I wrote a comparative review of the offerings more than 10 years ago. But these images are much more related to the shape, format and architecture of Paris Hilton than those of the Hilton in Paris. One of the most impressive novelties of the past few years is the Illustrata database which was launched this year. It uses deep indexing to provide metadata and thus search criteria for more than a million of illustrations extracted from 160 scholarly journals at the initial release.

I had a long review of this innovative database in Peter's Digital Reference Shelf column hosted by Gale, and I understand that the database is available in the exhibit hall. It shows you a thousand times better than what I could try to describe with words in the precious little time I still have. Take a test drive, to see that rarely is the adage so true that a picture (or chart or graph or table or cross-section) is worth a thousand words, and often much more. The superb FactSearch database show human-created fact-laden passages from source documents. Showing the facts provide immense help in cutting the infoglut. This database from Pierian Press never received appropriate recognition, and OCLC stopped updating the database. Luckily, it is still available directly from the publisher.

Linking to the digital items licensed/owned by the library

Open URL resolvers can much enhance the utility of the abstracting databases by checking if the user's library has a subscription for the volume and issue of a journal in which the abstracted article was published, and signaling this in the abstract record. Of course, you need to licence a link resolver, or do you?. Well, it depends. Even if you don't have a link resolver but have a subscription to a ProQuest database, its new OneClick feature will alleviate a common limitation in the availability of full text articles. Practically, all the commercial scholarly publishers apply some moratorium of 6 to 24 months which means that the most recent issues would not be available in a third party's full text database.

This is the case with ProQuest ABI/INFORM - which most of you probably know as its is an excellent and often unique source for information and library technology topics (although few users are aware of this). The British publisher, Emerald, for example, has a one year moratorium on *Online Information Review*. ProQuest has full text coverage of it, except for the past year. But here comes the pleasure of OneClick. A small symbol above the records for papers in the current issue will indicate that the user's library has a digital version of the paper through the library's subscription to the digital archives at the publisher, which does not have such moratorium.

On the fly enhancements for cherry picking the final results

In order to facilitate the selection of the most promising items from the hundreds of hits, the current relevance ranking algorithms of the largest aggregators would not suffice. In citation-enhanced databases the most obvious feature would be the sorting of the records by their citedness. Only Web of Science and Scopus offer that option, and the latter without regard to the size of the result set. This is a mighty feature. You can easily find out, for example who is or who was the most cited author of the Czech Republic, or of any topic, journals or institutions that you define by the search. It could only be better if a relative citedness count, the citedness per year would also be displayed and used as a sort criteria.

What would help significantly the cherry picking activity of the users is mashing bibliographic data in the catalog or in any of the databases with facts or even factoids gleaned on the fly from a) other databases licensed by the users' library and b) from open access databases.

The digital libraries of most publishers now do show for free the list of citing articles published in their own journals (when they recognize their own, which cannot be taken for granted as Wiley Interscience demonstrates it). For subscribers of JCR, records from WoS has a link to the journal's record in JCR. It may be even more efficient to show the journal's Impact Factor directly on the screen of the article record instead of taking the user away from the current database to the JCR database.

Similarly powerful mash-ups are already in place even for non-subscribers. For example, for many journals, their digital facilitators, HighWire Press and Atypon show the number of citations received by a given article in Web of Science. This is a highly informative clue. To get this information, the users' library does not need to have a subscription to WoS. The arrangement is made between Thomson ISI and the digital facilitator in a behind the scenes licensing transaction. To see the citing article records from WoS a subscription is needed. Such dips from Scopus will also likely to be available, providing another angle. Of course, there are also links to launch a search in the free Google Scholar to find the records in its collection. It did not work well, therefore HighWire Press removed this link, but retained the link for launching an author search in Google Scholar. I am somewhat relieved because sooner or later the next step would have been to fetch the citedness scores reported by Google Scholar, and these are often hugely inflated, nonsense scores as I discussed above, and demonstrated in my tutorial session yesterday.

Of course, there are several other development trends that will shape the face and not only the interface of academic and professional online services. I am testing now the variety of mash-ups that can provide excellent clues for the users who want to make an informed decision..

Suffice it to say in closing that every academic and professional online service should have a software that would at least try to a) understand mispronounced or misspelled words, b) make sense out of simplistic or garbled queries, c) guide the users through choosing the right databases, right search words, right synonyms, best qualifiers and filters for refining the search results, d) provide clues through adding novel and/or mashed-up facts, factoids, tidbits and snippets and e) facilitate the refinement of the query in an intuitive way or cherry-pick the ones from the final results most pertinent to the users.

Enjoy the conference, talk to the exhibitors, fellow librarians and other information professionals. I am one of them even if I just teach library and information science & technology and do not work anymore in a library.