péter's picks & pans

Péter Jacsó
University of Hawaii

Vivísimo, Central Search, *TIME* Magazine, and the Open Directory Project

The first pick is a powerhouse combo of a federated search engine enhanced by Vivísimo's clustering technology in the new release of Serials Solutions' Central Search. This will make most library patrons appreciate the quality and variety of information and the guidance on offerings from their libraries. The other pick is the *TIME* magazine archive with more than 320,000 articles covering all 84 years of the news magazine—a very convenient and reliable ready-reference source, free for subscribers of the print edition. The pan is the Open Directory Project, which has entries that have become so outdated and misinforming that it triggered the creation and use of the NOODP metatag to alert Web search engine crawlers to use the metadata of the site instead of the ODP description in their results lists.

> "I know that collaborative Web projects by volunteers are the craze and I admit their psychosocial advantage for members of the Lonely Hearts Club, but I prefer professionalism.

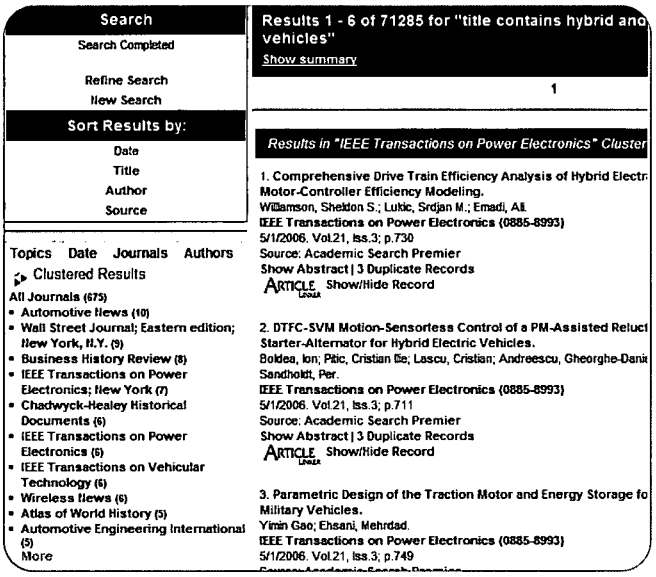
VIVÍSIMO AND CENTRAL SEARCH

Central Search is Serial Solutions' federated search engine—not the typical database I usually review in this column. I cover it here because it brings the best out of the mix of the abstracting and indexing (A&I) and full-text databases, journal archives, digital encyclopedias, and ready-reference sources to which librarians subscribe or—in cases of open access sources—provide access to. Through its Article Finder link resolver, Central Search also enhances the results lists of A&I records with links to the primary documents—which are under the patrons' noses, but people remain oblivious to them. Central Search was a smart and pretty federated search engine. Since its enhancement by Vivísimo's clustering engine in October, it has become a very smart and very pretty resource discovery tool.

Many other information services have been doing result clustering to provide at-a-glance clues about the composition of the results list (typically from a single database) by topics, keywords, source publications, authors, document types, or ages. However, it's not to the tune of hundreds of databases from different services, and it's typically by a single criterion rather than by a number of data elements at one fell swoop.

The entries in the clusters (along with the information posted) provide hints for drastically and instantly reducing the results set to a manageable ad hoc temporary subset by user-chosen topical criteria or other angles, such as journals or authors listed in the cluster sidebar.

The entries, being from multiple sources, can be redundant and inconsistent. History shows that they are often that way even within a single database. The heterogeneous mix of databases from different content providers increases the inconsistency. The cluster provides an indirect warning on what a Sisyphean task it would be to do a cross-database search by descriptors, authors, or journal names—and it offers a bypass solution to the problem. It also gives a sense of the overlap among databases.



Cluster options for the results list in Central Search

If check boxes were displayed instead of the square bullet gizmos, users could consolidate the spelling variants of journal names such as the ones for *IEEE Transactions on Power Electronics*. For the journal cluster, Serials Solutions is in the best position, automatically and instantly, to do much of the consolidation using the many journal name variants in its A–Z source list (which is more like a knowledgebase). There are no "InstaSolutions" for the author names and descriptor variants scattered around in the cluster list, but it is an acceptable compromise in exchange for the immense power offered even by the not exactly squeaky-clean cluster terms.

This is exactly the type of application that would make users of the general Web search engines (which serve up increasingly polluted results from an unpredictable medley of sources) appreciate their libraries and understand that the most efficient way of finding pertinent information is to use the professional databases available with their free library cards. For libraries, the yearly cost for 100 databases with a consortial discount is less than $200 a pop. The Vivísimo enhancement comes as a bonus and certainly will ensure much better use of databases—each of which costs several thousands of dollars, yet they often remain unknown and unused. This is a big bang for the buck.

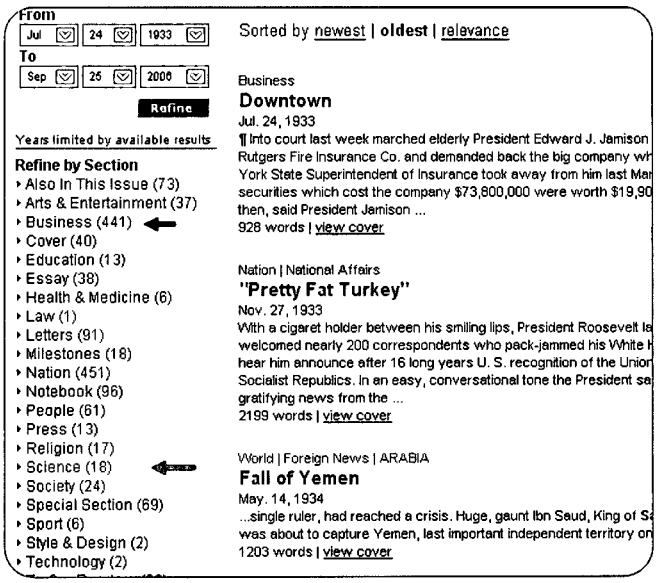
TIME ARCHIVE

Of the three largest news magazines, *TIME* is far ahead of *U.S. News & World Report* and *Newsweek* in spoiling its print subscribers with access to a digital archive. I often use it as a ready-reference source to get reliable, contemporaneous news about an event or a person of the pre-1960s era that is not covered by other databases of general interest magazine and newspaper articles (except for ProQuest's Historical Newspapers database family). The entire *TIME* archive from 1923 is available free of charge for current subscribers. This represents a 320,000-item full-text collection. Half the collection consists of short articles (perfect for ready reference) and, for those with a longer attention span, the other half is a mix of medium and long articles, including nearly 3,800 cover stories.

There is an impressive gallery of 4,370 cover pages, perfectly browsable in a given time frame selected by the user. There are predefined subsets of thematic cover pages about—among others—musicians and music industry-related personalities (134), China (61), religion (169), presidents (297), and the Persons of the Year (80). There are other lists of cover pages, allowing users to create their own minigalleries by keyword combinations. You could, for example, create a minigallery of cover pages about, say, Iran (38), global warming (9), or terrorism (68). There are also predefined special collections about historical events, arts and entertainment, health, the economy, social issues, sports, and their subcategories such as cancer, the bird flu, and cloning.

The number of direct search options is very limited. There is phrase searching (the default operation is AND), which limits results to cover stories, cover pages, and time periods. However, the results set is automatically clustered by sections of the magazine (both the current and old sections), authors, and document lengths, which would orient users zooming in on a special subset.

Some of the cluster terms and their postings may look strange for the search topic. The query `Saudi` yields the most hits in the sections about terrorism and business, but religion and science appear 17 and 18 times, respectively. The association of the latter term with the subject of the query seems very questionable. Indeed, most of the articles are not about scientific developments in Saudi Arabia. The search



Results list with clustered section terms (and then some) from the TIME *archive*

term just appears in passing, for example, in the article about space tourism where a Saudi prince was among the first five candidates who would have paid for the privilege of being a civilian passenger. Lacking descriptors, there should be an option to use the equivalent of the ATLEAST command of LexisNexis (or the FREQ command of Ovid) to limit the full-text search to those articles where the query term appears at least a specified number of times.

OPEN DIRECTORY PROJECT

Unfortunately, the largest and most widely used Webwide directory is now full of inferior digital resources and omits many of the best resources. Originally created to guide users to the most valuable Web resources, the sites identified by the directories were selected by a few competent individuals who evaluated the sites (or at least took a closer look at them) before an entry was added. The popularity of Yahoo! in the early years of the Web was due to its well-structured directory, which acted like a Michelin guide. Not accidentally, one of the first employees hired by Yahoo! was a librarian. For many years, Yahoo! did not have and did not need a search engine.

Then came the ODP, the Open Directory Project (http://dmoz.org), and the senseless race to have the largest directory of Web sites started. As of early September 2006, there were more than 4 million sites spread across more than 590,000 categories created by 74,126 editors in the "largest human-edited directory of the Web." I know that collaborative Web projects by volunteers are the craze and I admit their psychosocial advantage for members of the Lonely Hearts Club, but I prefer professionalism. The ODP also functions as the directory for Google (and hundreds of other Web portals and sites), so it has a far wider reach than any other directory. That makes its weaknesses more of a problem.

The ODP is the best (or rather, the worst) example to illustrate that quantity does not quality make. ODP editors (and ODP users) get lost in the jungle of categories and subcategories. Hundreds of low quality and irrelevant sites fill the subcategories while the best sites are ignored. Even the highly selective directories, such as the Librarians' Internet Index or the reinvigorated Internet Public Library, have entries for a few questionable quality, open access, ready-reference sources and miss a few important ones. However, the high ratio of excellent to poor entries still makes them a good choice.

A look at the top level of the Encyclopedia category in ODP sheds some light on the problem. The first entry refers to concise entries. The ODP editors have not realized that the Columbia Concise edition was replaced by the Columbia Encyclopedia 6th edition a long time ago—and the site is HighBeam. The information about the Bartleby.com site describes it as the online edition of the Columbia Encyclopedia. It is far, far more than that. The Encyberpedia site does not have more than 10,000 links to the best reference sites on the Internet; I doubt there is any beyond the topical domain of Scripophily. The Encyclopaedia Britannica site offers the entire text only for subscribers. There is a Concise Britannica

- Encyclopedia.com from Electric Library ☆ - Free resource offers more than 50,000 concise articles.
- Bartleby.com - Online publication of the Columbia Encyclopedia, Sixth Edition.
- Encyberpedia - Electronic encyclopedia with its own content as well as over 10,000 links to the best reference sites on the Internet.
- Encyclopaedia Britannica Online - Provides complete text of Encyclopaedia Britannica with search capabilities, related links, and multimedia enhancements. Free trial.
- 1911 Encyclopedia Britannica - Offers the 11th Edition Encyclopedia Britannica which is now in the public domain. All articles are searchable and alphabetized.
- Encyclopedia FunTrivia - Thousands of categorized entries of trivia - an invaluable resource for answer seekers.
- Encyclopedia Smithsonian - Links to on-line Smithsonian resources and answers to frequently asked questions from A to Z.
- EncycloZine - Portal/concise encyclopedia offers fun activities for learning and exercising the mind. Daily articles introduce new topic in the arts, humanities, science, and technology, and provide links to related sites.
- Free Internet Encyclopedia - An encyclopedia composed of information available on the Internet. MacroReference contains references to large areas of knowledge, FAQs where available, and pointers to relevant areas of the MicroReference. MicroReference contains short bits of information and references to specific subjects.
- How Stuff Works - Articles explain how hundreds of things work, from elements of the human body to space-age technology.
- Informationsphere.com - Free online encyclopedia that allows users to submit their own articles for inclusion in its database.
- The LoveToKnow Free Online Encyclopedia - Browsable. Based on the eleventh edition of the Encyclopedia Britannica, first published in 1911.
- Nuttall Encyclopædia of General Knowledge - Approximately 14,000 entries on topics including art, geography, literature, philosophy, religion. Originally published in 1907.
- The Probert Encyclopaedia - Reference source with several hypertext linked entries covering a broad variety of subjects.
- Wikinfo - User edited encyclopedia based on Wiki software. Contains articles, information on editing, and a list of differences between this encyclopedia and Wikipedia.
- Wikipedia - A free encyclopedia built collaboratively using Wiki software. (GNU Free Documentation License)
- World Book Encyclopedia - Online version of the complete reference work along with dictionary, atlas, links, magazines, historical documents, audio, video, images, and 3D photographs. Requires free temporary registration with option to buy continued access.

The Encyclopedias subcategory from the Open Directory Project

site, with its entire text in open access, but this source does not appear in the ODP. The two entries about the 1911 edition of the Encyclopaedia Britannica do not compensate for this omission. Encyclopedia Fun Trivia does not really belong to the encyclopedia category; it is a quiz collection. Neither does Encyclozine, which is a potpourri of links to Web sites. The Free Internet Encyclopedia (which gets the longest entry) is a pathetic site of links (most of which are dead), with a selection of entry terms lacking rhyme or reason. Informationsphere is yet another pathetic site, which appears to have been created for the sole purpose of satisfying its developer. The World Book encyclopedia has a fine concise edition that does not require registration.

While there are many junk reference sites in the Reference category (and elsewhere), the blue chip sites of the genre are not included at all in the ODP (let alone in the right category). Examples include the Encarta Encyclopedia (concise edition), the Concise Hutchinson Encyclopedia, the beautiful and informative Encyclopedia of New Zealand, the Canada Encyclopedia (both the English and French editions), its junior edition, the very current Crystal Reference Encyclopedia (created by the guru of encyclopedias, David Crystal), or the concise edition of McGraw-Hill Encyclopedia of Science and Technology—just to name a few of the best open access encyclopedias.

The majority of the entries in the ODP are so outdated and misinformed that recently Google and MSN decided to support the robot tag NOODP so Webmasters can avoid the use of the ODP entries for sites in their results lists. Google also demoted its version of the ODP, putting it on a back burner from the front page. Better later than never.

Péter Jacsó *(jacso@hawaii.edu) is professor of library & information science at the University of Hawaii's department of information and computer sciences.*

Comments? Email letters to the editor to marydee@xmission.com.