péter's picks & pans

Péter Jacsó
University of Hawaii

OECD.Stat, Astrophysics Data System, Wiley-Blackwell

> The picks include the new, promising OECD.Stat and the splendidly enhanced Astrophysics Data System (ADS).

The picks include the new, promising OECD.Stat, which functions not merely as a portal but as a sophisticated, state-of-the-art federated search engine for the many statistical data sets created by the Organisation for Economic Co-operation & Development. My second pick is the splendidly enhanced Astrophysics Data System (ADS), one of the pioneers of open access bibliographic databases. ADS covers all branches of physics, not just astrophysics as its name suggests. It is a model of how to develop and enhance, with great competence, traditional databases through cited references.

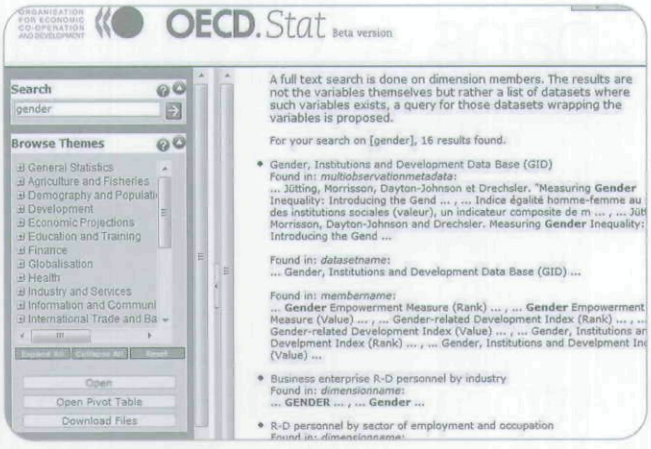
The pan is the very unpromising Wiley-Blackwell database with the serious, long-standing software deficiencies of Wiley InterScience that will also affect Synergy, the far better digital library of Blackwell. Synergy is to cease operations in July 2008—unless librarians mount a campaign to retain it in order to avoid the chaos that they have witnessed with their print subscription since the acquisition of Blackwell by Wiley last year. There should be a moratorium, at least until Wiley-Blackwell releases a well-tested, state-of-the-art service, which it promises to do only in 2009.

the picks

OECD.STAT

I tested the free beta version of OECD.Stat (http://stats.oecd.org/wbos), which is destined to replace the current browser (Beyond 20/20) for SourceOECD in 2009. Some functions were not yet operational, but the key functions were there. It is interesting and encouraging to see the extension of the concept of federated searching from the bibliographic sphere to the factographic area. The United Nations and World Bank have been doing it, and now OECD is following suit.

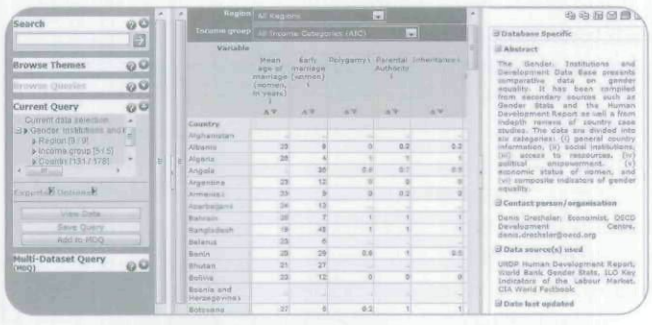
OECD.Stat allows searching across 50 statistical databases compiled by OECD. They cover the traditional demographic, economic, labor, health, educational indicators, and then some. You can browse through categories and subcategories (called themes) by smoothly expanding and collapsing them, or just entering a query such as gender to see what kind of statistics are available. Actually, for this query there are 12 hits. One is from a new database dedicated to Gender, Institutions and Development (GID-DB); the others are statistical data sets that have gender-specific variables.



OECD.Stat search results for a search on **gender**

There are statistics for 178 countries and five income groups. Of course, not all the indicators are available for all the countries, although both the choices for countries and income groups include "all of the above" groupings. The tables can be limited by these aspects, and also by eight major predefined regions. The content of the tables are dynamically generated, and the customization is very good. With that said, it would be better if the regions could be saved and customized by users to apply them repeatedly, such as limiting a group to predominantly Islamic countries.

The GID-DB excels not only for its options but also for the specific indicators, which go beyond the usual gender statistics (maternal mortality, marital status). Most of the special indicators—unusually, but understandably—are given on a scale from 0 to 1. The criteria include traits such as the acceptance of polygamy (from not accepted to entirely accepted); the share of parental authority between mother and father (0=equal, 1=father only); the levels of inheritance practices favoring male heirs (0=no preference, 1=only males can inherit); the existence of laws against rape, sexual assault, domestic violence, and sexual harassment (0=specific legislation in place, 0.50=being drafted, 1=none); women's right to own property (0=full,1=no); and obligation to wear a veil in public (0=no, 1=mandatory).



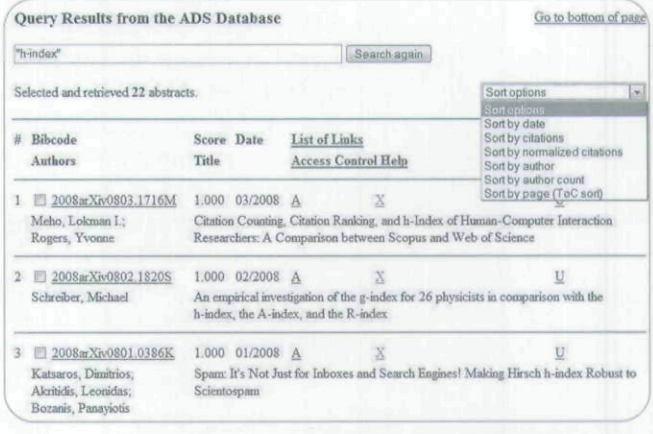
Unusual but highly informative gender equality indicators

The screen is split into three panes, but the left and right panes can be hidden to display more indicators. Each column can be sorted. The tables can be immediately emailed. The results can be downloaded into a spreadsheet or a text file. Access to the complete set of databases will not remain free, but a subset is likely to be offered without subscription. [*OECD.Stat is still in beta, so it can change without warning. For example, when Péter reviewed it, the search box was in the left pane. The latest version puts it at the top right-hand corner of the page. These types of incremental changes may or may not be permanent.* —Ed.]

ASTROPHYSICS DATA SYSTEM

The Astrophysics Data System (ADS) has been around for more than 15 years. It grew from an indexing/abstracting database into a very sophisticated, highly customizable bibliographic-factographic mega-database of full-text-searchable, cited-reference-enhanced digital collection scientific publications (www.adsabs.harvard.edu). In spite of its name, it covers a far broader spectrum than astrophysics. At the end of March 2008, it has 6.7 million records, with roughly 25% related to Astronomy and Astrophysics, nearly 70% to Physics and Geophysics, and the rest to miscellaneous fields of sciences.

In the early years the full-text papers were not searchable because they were scanned as images. However, in the past few years a large scale OCRing project converted the image text to plain ASCII text, which has hugely enhanced the recall of searches, particularly in cases where the query included terms that did not appear in the title or the abstract or in the cited references. Of course, OCRing is (to put it mildly) less than 100% accurate, which is not a problem for terms that appear repeatedly in the body of the text, but it can be lethal for linking to cited references and counting citations. Enter the professionalism of the small group of the ADS developers, who did an outstanding job in correcting erroneous cited references to match those with items available as master records in ADS.



Excerpt from a result list with a variety of coded links to the abstract, full text and/or image version of the preprint or reprint versions of the papers, and its cited and citing references

This in turn led them to develop an impressive variety of citation counts with and without self citations, as absolute and normalized values (divided by the number of co-authors), from refereed and nonrefereed sources. Results can be sorted, among many other options, by absolute and normalized citation counts. Not even Web of Science or Scopus offer such a variety of citation-based options. Most other databases don't have cited reference enhanced records at all, let alone such a variety of citation counts that can be used for ranking results by frequency of references as a clue for importance.

Most databases still rely merely on free text and controlled vocabulary searching and highly questionable relevance ranking when the web—which itself is based on links, the high-tech equivalents of references—is already deep in its teens. I wish ADS would also offer a relative citation count to reflect the number of years the paper has been citable.

The software has other unique or rare features, such as searching authors as first (or only) author (an important trait in academia for some), or left-handed truncation (along with optional stemming) to retrieve, for example, symmetry, symmetric, asymmetry, and supersymmetry just by searching `*symmetry`. The extent of customization and personalization of the search and output options is also impressive, along with the myADS current awareness service. This open access service is a prime example for cooperation among scientists, programmers, and librarians—even if you don't know much about astrophysics, which is the nice way of putting *my* embarrassing lack of knowledge of physics.

There is a reason that members of the team (including a librarian) are identified on the front page of ADS (except for Günther Eichhorn, who recently left for Springer and won the Royal Astronomical Society's Award for Services to Astronomy). They have done an excellent job in improving citation matching (which is unscholarly and flimsy in Google Scholar and simply pathetic as a just-introduced feature in Windows Live Academic). ADS developers implemented many pioneering features for real-life use and published several papers that give good ideas for others who need to develop state-of-the-art back-end applications for scientific databases, not just pretty interfaces. ADS is to astronomy, astrophysics, geophysics, and many other *physics what PubMed with PubMed Central combined are for medicine.

The ADS personnel are:

- Dr. Stephen S. Murray - Principal Investigator - *ssm at cfa.harvard.edu*
- Dr. Alberto Accomazzi - Project Manager - *aaccomazzi at cfa.harvard.edu*
- Dr. Michael J. Kurtz - Scientist - *mkurtz at cfa.harvard.edu*
- Carolyn Stern Grant - Programmer - *stern at cfa.harvard.edu*
- Edwin Henneken - Programmer - *ehenneken at cfa.harvard.edu*
- Elizabeth Bohlen - Computer Specialist - *ebohlen at cfa.harvard.edu*
- Donna Thompson - Library Specialist - *dthompson at cfa.harvard.edu*

The members of the team are listed on the ADS homepage for good reason.

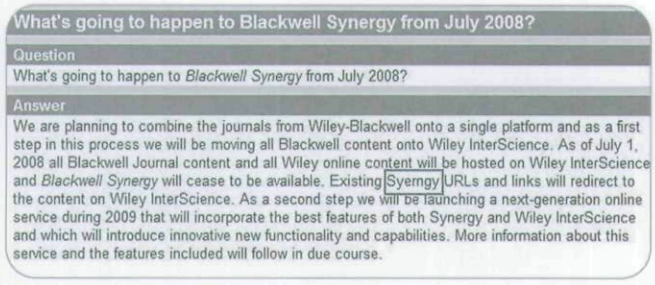
ADS developers implemented many pioneering features for real-life use and published several papers that give good ideas for others who need to develop state-of-the-art back-end applications for scientific databases, not just pretty interfaces.

the pan

WILEY-BLACKWELL DIGITAL LIBRARY

A year ago SpringerLink and Blackwell Synergy were my picks and Wiley InterScience digital service (www.interscience.wiley.com) was my pan. SpringerLink is unlikely to have problems, as it hired away Günther Eichhorn from ADS to be the director of its abstracting/indexing services. *He* certainly knows how to develop, maintain, and run cited-reference-enhanced database services very well. But I am very worried about Blackwell's Synergy being subsumed into InterScience. I wrote a year ago that "I can only hope the acquisition of Blackwell by Wiley will not mean the incorporation (and thus the partial incapacitation of the features) of the digital collection of the former into the poorly performing Wiley InterScience collection." Now I have to give up that hope; I have seen the official decision that all Blackwell content will be moving onto Wiley InterScience and Blackwell Synergy will cease to be available.

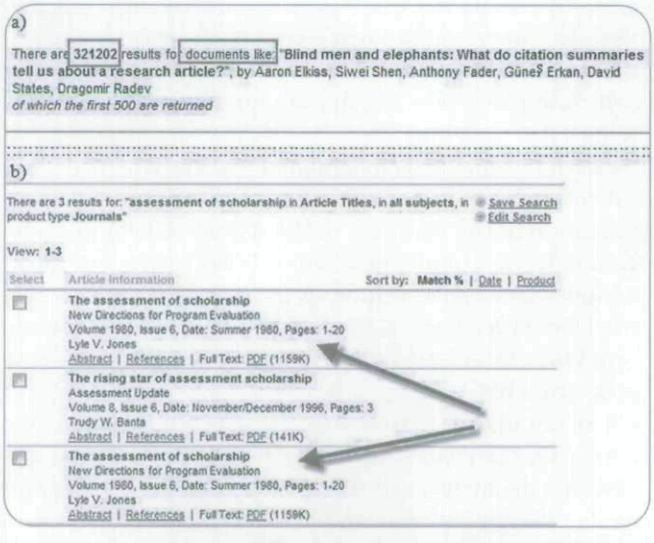
It is not a good sign that Synergy is misspelled in the statement, but chalk it up to the disappointment and distress that the merger and this decision caused. Be ready for similar disappointment and distress for a long time as Wiley-Blackwell "will be launching a next generation online service during 2009." I think they've put the cart before the horse. Even if this next generation service would ever work



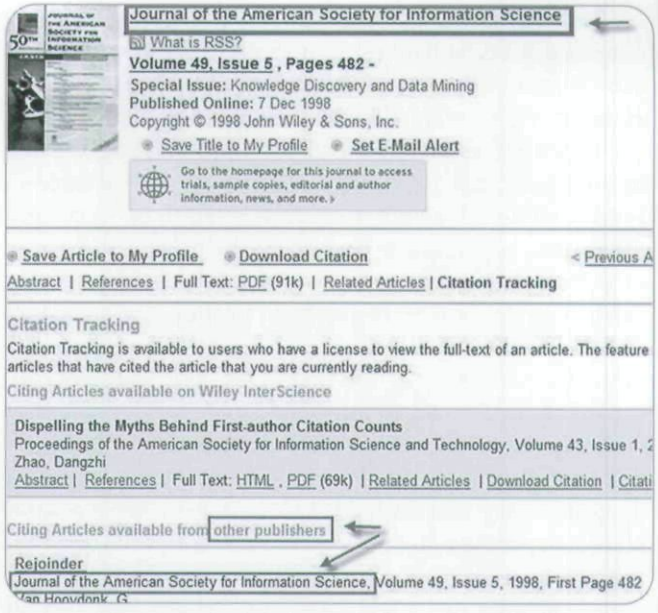
The synergy will be as good as the spelling of "synergy" in this statement.

nearly as well as Synergy, it would have been the minimum of precautions to wait until this is proven and tested by independent and competent experts, not just declared by the management of Wiley-Blackwell.

It should have been done the other way around. Wiley has never had the ability to develop a sophisticated online journal service, nor has it cared about fixing its deficiencies (www.jacso.info/PDFs/jacso-spingerlink-blackwell-wiley.pdf). Beyond the software issues mentioned there, InterScience goes into its "Forrest Gump" mode when trying to find articles related to the one being shown. The numbers are far more than a tad excessive, even if we discount the unknown number of duplicates that came up for another query. The set of 321,202 records for documents which are purportedly like the *JASIST* paper titled "Blind men and elephants" (which discusses how differently authors describe the essence of their cited references, and *not* the trials and tribulations of sightless people confronting a pachyderm), brings up an insanely large set of "related" records. It apparently does a keyword search with



Excessive number of items allegedly related to an information science paper (a), even when discounting it for possible duplicates found for another search (b).



InterScience denies ownership of JASIS *for an article which is cited by another item in—well,* JASIS.

a Boolean OR, and brings up all records that include the words: `elephant or blind or men`.

It is even more disconcerting that the Alzheimer's condition of Wiley InterScience further deteriorated after my last review. In its citation tracking of an item published in *JASIS*, it identifies a citing item from *JASIS* under the category of "citing articles from other publishers." This is worse than a denial in a paternity case for a child who is the spitting image of the defendant.

I am sure the manager in charge of InterScience and the database merger will get a pat or two on the shoulder and praise for "doing a heck of a good job" by cutting costs, after the subsumption of Synergy into InterScience, while librarians and their patrons will be shortchanged by the cheap solution. The problems of this acquisition popped up in the frustrations of the renewal of journals for 2008. To get a feel for the mess, run this query in Google: `Blackwell-Wiley problems serialst` (yes, serialst is the correct spelling!), and then follow the discussion that began when I started to work on this column. For the digital archive merger, it is likely to get much worse. The ALA conference may give the best chance for serials (and reference) librarians to voice their concern about the plan to take effect in July, before Synergy is brought down in more ways than one.

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

Comments? Send email to the editor (marydee@xmission.com).