

## Savvy searching

### Searching “unsearchable” open access digital collections

Savvy searchers know that just because the open access (free) archive of a journal lacks a search engine, this does not necessarily mean that they have to put up with browsing the issues one by one, scanning the table of contents, sometimes the abstracts, and peeking into promising articles to see if they are relevant for their topic. Odd as it may seem, there is often a way to search open access archives that are not yet endowed with native search capabilities, or have very poor search engines.

### Open access digital collections with search engines

Open access archives and other collections which do not require subscription access keep growing on the Web. The most well-known are the e-journals, but there are many archives of conference papers and dissertations, as well as a growing number of preprint collections. Many of them are very valuable, describing current research and published a short time after submission, even when they go through the refereeing process.

The best of these archives have good or excellent search engines, based mostly on full-text keyword searches. Some of them use customised versions of commercial search engines designed for home-grown archives. One of the best of them is Atomz, which is used by more than 48,000 sites to allow users to search, for example, the biographies of the San Francisco Ballet, or information files of the Massachusetts Motor Vehicle Department, or articles of the British e-journal, *Information Research* (<http://informationr.net/ir/search.html>) (see Figure 1), or the American e-journal, *Issues in Science & Technology Libraries*.

Atomz (see Figure 2) is an interesting solution, as publishers do not need to worry about setting up software; their sites are remotely searched by Atomz, whose specialists analyse the Web site, then negotiate the personalisation issues of the query template and the output options to customise the behaviour of the search engine. This is capable software which allows limiting of the searches to specific elements of the

HTML pages (title, descriptors, etc.), and even allows Soundex searching that is very convenient when you want to search for articles about, say, information-seeking behaviour and you do not need to bother with the spelling variations of behavior/behaviour.

Other archives and collections use search engines that are developed in-house or licensed for installation on the server where the archives are stored. The best examples for this include such outstanding information science and technology e-journals as *First Monday* and the *Journal of Electronic Publishing*. Some sites have search engines that are not nearly as capable, or have not been tuned to the special characteristics of the digital archives of publications.

### Open access digital collections without search engines

The first priority of the publishers of electronic journals and other document archives is to build the collection. There is not much reason to spend time and money on implementing a search engine when there are only a few articles that can be easily browsed. This is probably the reason why the publisher of *CyberMetrics* (<http://www.cindoc.csic.es/>) does not yet offer a search engine to the small but very interesting collection of a dozen articles. (For disclosure: I am on the advisory board of another, Spanish-language journal published by CINDOC.)

We are willing to browse through individual issues to see if there is an article or conference paper or PowerPoint file in a small digital collection, clicking on each issues, scanning the table of contents, then the most promising document(s). In doing so we learn more about the context, and can bump into other articles which are of possible interest in another project. Also, if the open access electronic journal or conference proceedings are abstracted/indexed by one of the secondary databases, we can go straight to the article, knowing its numerical-chronological sequence. Then there are the hotlinks from other documents to a specific HTML document in the archive that takes us directly to the paper. Unfortunately, in real life the savviest searchers often do not have the luxury of time, and want to cut through the clutter to find highly relevant articles in the shortest possible time. That is when searching the “unsearchable” archive becomes a necessity. One of my solutions is to use Google to locate papers in open access digital archives.

Figure 1 Atomz query template for the digital archive of *Information Research*, offering several search options

**Information Research**  
an international electronic journal

**Full-text Search of the contents of the electronic journal**  
*Information Research*

To search the whole Web try **Google**

---

Enter keywords in the box below  
Use the buttons below to specify the kind of search you want. In particular, check the **Exact phrase** button if you want to avoid getting too many irrelevant hits. For further advice on searching, check out the [Search Tips](#) page.

Search for:

Match:  Any word  All words  Exact phrase  
 Sound-alike matching

Within:

Show:  results with

Sort by:

Figure 2 Atomz search results from the digital archive of *Information Research*

SEARCH RESULTS 1 - 7 of 7 total results for **information seeking behavior**

Powered by Atomz [Sort By Date](#) | [Show Summaries](#)

1. **A STUDY OF THE DEVELOPMENT OF THE DIGITAL RANCH ...**  
64% <http://www.shel.ac.uk/~is/publications/intres/paper16.html>
2. **Reference group theory with implications for information studies: a theoretical essay ..**  
54% <http://www.shel.ac.uk/~is/publications/intres/6-3/>
3. **Facilitating access to and use of bioinformatics resources ..**  
51% <http://www.shel.ac.uk/~is/publications/intres/6-2/ws1.html>
4. **University students' information seeking behaviour in a changing learning environment ..**  
51% <http://www.shel.ac.uk/~is/publications/intres/isic/>
5. **Paranormal information seeking in everyday life .. part II: the paranormal in information action ..**  
51% <http://www.shel.ac.uk/~is/publications/intres/isic/kan>
6. **The use of certainty and the role of topic and comment in interpersonal information seeking interaction ..**  
51% <http://www.shel.ac.uk/~is/publications/intres/isic/yoou>
7. **Facilitating access to and use of bioinformatics resources ..**  
51% <http://www.shel.ac.uk/~is/publications/intres/6-2/ws1a.html>

### "Googled" searching of open access digital collections

Google is the most popular search engine for many reasons. It has one of the largest collections of documents published on the Web. It not only has indexes and pointers to those documents but very often an archived (cached) version of them. It indexes not only HTML documents but also PDF, Word, Excel, PowerPoint, and RTF files. It has an appealingly simple interface and a spectacularly smart search engine that usually brings up the most relevant items

first, or among the first dozen or so results. It can also offer correctly or alternatively spelled versions of words and phrases in case you misspelled your query. Not even professional search services can do that, although there are a few exceptions such as Lexis-Nexis, which is good at recognising variant spellings.

From the perspective of searching "unsearchable" digital collections, the advanced mode of Google is the most important feature of this splendid software. Remember, it can do wonders in no time because, when crawling the Web, it collects and saves the documents, translates them – if needed – into HTML format (albeit not always perfectly) and then indexes the documents. When running an advanced search, I make Google search this harvested collection on its own server, not on the server of the publishers. The trick is to use – in addition to the query term – in the advanced search template the domain of the site that hosts the collection. This is often a partial URL, which includes the name or acronym of the journal. For example, when searching for articles about impact factor in the archive of the *Journal of Medical Internet Research*, you specify in the exact phrase cell of the search template the term impact factor (without quotes), and <http://www.jmir.org> in the domain cell (Figure 3). Google finds in 0.04 seconds seven records in its own collection that match the query. This does not necessarily mean seven separate articles, but seven HTML files because an article may consist of, say, two interlinked HTML files, one for the text, and another for the figures (see Figure 4).

Figure 3 "Googled" query in the "unsearchable" digital archive of the *JMIR*

Advanced Search Tips | All About Google

**Google** **Advanced Search**

Find results with **all** of the words  10 results

with the **exact phrase**

with **any** of the words

**without** the words

Language Return pages written in

File Format  return results of the file format

Date Return web pages updated in the

Occurrences Return results where my terms occur

Domains  return results from the site or domain  [More info](#)

SafeSearch  No filtering  Filter using SafeSearch

Figure 4 Results for the exact phrase query term impact factor from JMIR

Searched pages from [www.jmir.org](http://www.jmir.org) for "impact factor".

JMIR - Hernandez - Tabel 2  
Table 2. Correlations among the number of daily visits to the web sites, the **impact factor** of their authors or editors, the grade of update, and the number of ...  
[www.jmir.org/1999/1/e1/table2.htm](http://www.jmir.org/1999/1/e1/table2.htm) - 3k - [Cached](#) - [Similar pages](#)

JMIR - Hernandez - Tabel 3  
... 3. Correlation among the number of links and visits to the web sites, the **impact factor** of their authors, and the time since the last update, and the results ...  
[www.jmir.org/1999/1/e1/table3.htm](http://www.jmir.org/1999/1/e1/table3.htm) - 5k - [Cached](#) - [Similar pages](#)

JMIR - Hernandez - Tabel 4  
... inbound links. The weeks since the last update, the number of daily visits to the web sites and their editor/author's **impact factor** are also provided. In ...  
[www.jmir.org/1999/1/e1/table4.htm](http://www.jmir.org/1999/1/e1/table4.htm) - 35k - [Cached](#) - [Similar pages](#)

J Med Internet Res 2000 Vol. 2 Iss.1 Suppl. 2 - ...  
... the accuracy of responses. In three studies authors used the cumulative **impact factor** of the published research of the mailing list contributors as an ...  
[www.jmir.org/2000/3/suppl2/e9/](http://www.jmir.org/2000/3/suppl2/e9/) - 22k - [Cached](#) - [Similar pages](#)

Journal of Medical Internet Research 2000 - Practical...  
... to a specific, automatically calculated personal value to be added to something similar to a personal **impact factor** (which would derive from the comments). ...  
[www.jmir.org/2000/3/e15/](http://www.jmir.org/2000/3/e15/) - 14k - [Cached](#) - [Similar pages](#)

Journal of Medical Internet Research 1999  
... of daily visits to those websites, the time since their last update, the **impact factor** of their authors or editors, and the number of websites linked to them ...  
[www.jmir.org/1999/1/e1/](http://www.jmir.org/1999/1/e1/) - 48k - [Cached](#) - [Similar pages](#)

Journal of Medical Internet Research 1999  
... journals are cited. There is also a measure of effectiveness, the **impact factor**, which normalizes the citations received by the selected journals and looks ...  
[www.jmir.org/1999/1/e4/](http://www.jmir.org/1999/1/e4/) - 31k - [Cached](#) - [Similar pages](#)

Using the domain name part of a URL may not always be distinctive. It may represent the home site of a university, or a publisher, but not that of the specific journal in which you are interested. This may yield some irrelevant results. Can you specify also the distinctive part of the URL in the domain cell? Not in Google, unfortunately. Consider, for example, the electronic journal, *Digital Technology & Law Journal (DTLJ)*, published by Murdoch University, in Australia. If you can specify only [murdoch.edu.au](http://murdoch.edu.au) in the domain cell of the search template along with your subject search term, Google will retrieve articles also from *eLaw*, the other electronic journal of Murdoch University (Figure 5), and possibly from other pages within the domain, but unrelated to *DTLJ*. In this example the first five items are from *eLaw*.

This is not necessarily bad, as you may find relevant articles in the other journal(s) too. I discovered the other journal of Murdoch University by being able to specify only the domain and the query term, fair use (see Figure 6). While *eLaw* is as useful as *DTLJ*, if you want to restrict the search to the latter, you should enter more of the URL, such as [murdoch.edu.au/dtlj](http://murdoch.edu.au/dtlj). The problem is that you receive no hits because the

Figure 5 Result for the exact phrase query term from two journals of Murdoch University

Searched pages from [murdoch.edu.au](http://murdoch.edu.au) for "fair use".

E Law: NetWatch: July 1996  
... Law course student papers, SUNY Buffalo School of Law (USA); Copyright & **Fair Use** STANFORD--Stanford University Libraries & Academic Information Resources, in ...  
[www.murdoch.edu.au/elaw/issues/v3n2/netw32.html](http://www.murdoch.edu.au/elaw/issues/v3n2/netw32.html) - 13k - [Cached](#) - [Similar pages](#)

[www.murdoch.edu.au/elaw/issues/v3n2/netw32.html](http://www.murdoch.edu.au/elaw/issues/v3n2/netw32.html)  
... <http://www.acsu.buffalo.edu/~hlmeyer/Complaw/complaw.html> Copyright & **Fair Use** STANFORD--Stanford University Libraries & Academic Information Resources, in ...  
8k - [Cached](#) - [Similar pages](#)

[www.murdoch.edu.au/elaw/issues/v4n4/netw44.html](http://www.murdoch.edu.au/elaw/issues/v4n4/netw44.html)  
... [www2.waikato.ac.nz/lawlib/decisions/menu.html](http://www2.waikato.ac.nz/lawlib/decisions/menu.html) Intellectual Property Copyright and **Fair Use** from Stanford University Libraries URL: <http://fairuse.stanford.edu> ...  
10k - [Cached](#) - [Similar pages](#)

E Law: NetWatch: December 1997  
... Intellectual Property. Copyright and **Fair Use** from Stanford University Libraries URL: <http://fairuse.stanford.edu/> International Law. ...  
[www.murdoch.edu.au/elaw/issues/v4n4/netw44.html](http://www.murdoch.edu.au/elaw/issues/v4n4/netw44.html) - 19k - [Cached](#) - [Similar pages](#)

E Law: Loose Strands in the Web: Meta Sites, Intellectual ...  
... [2] D Phan "Will **Fair Use** Function on the Internet" (1998) 98 Columbia Law Review 169 at 191. [3 ...  
[www.murdoch.edu.au/elaw/issues/v8n1/fones81\\_notes.html](http://www.murdoch.edu.au/elaw/issues/v8n1/fones81_notes.html) - 42k - [Cached](#) - [Similar pages](#)

Digital Technology Law Journal  
... it will be recalled that Part 1 recommended that Australia introduce a general **'fair use'** type of exemption along the lines of that contained in Section 106 of ...  
[www.law.murdoch.edu.au/dtlj/1999/vol1\\_2/bentley.html](http://www.law.murdoch.edu.au/dtlj/1999/vol1_2/bentley.html) - 29k - [Cached](#) - [Similar pages](#)

Figure 6 Result for exact phrase query term from one journal only

Searched pages from [murdoch.edu.au](http://murdoch.edu.au) for "fair use" dtlj

[PDF] [Digital Technology Law Journal](#)  
 File Format: PDF/Adobe Acrobat - [View as HTML](#)  
 ... 1 Number 2 [http://www.law.murdoch.edu.au/dtlj/1999/vol1\\_2/bentley.pdf](http://www.law.murdoch.edu.au/dtlj/1999/vol1_2/bentley.pdf) CLRC Report  
 Part 1 that Australia adopt a **fair use** defence and at Coolongatta through the ...  
[www.law.murdoch.edu.au/dtlj/1999/vol1\\_2/bentley.pdf](http://www.law.murdoch.edu.au/dtlj/1999/vol1_2/bentley.pdf) - [Similar pages](#)

**DTLJ: Regulating Speech on the Internet**  
 ... that this might stifle access to speech on the Internet is the **fair use** defence  
 which is available for educational and critical uses of Internet ...  
[www.law.murdoch.edu.au/dtlj/1999/vol1\\_1/blakeney.htm](http://www.law.murdoch.edu.au/dtlj/1999/vol1_1/blakeney.htm) - 75k - [Cached](#) - [Similar pages](#)

[PDF] **DTLJ: Regulating Speech on the Internet**  
 File Format: PDF/Adobe Acrobat - [View as HTML](#)  
 ... 1 Number 1 [http://www.law.murdoch.edu.au/dtlj/1999/vol1\\_1/blakeney.pdf](http://www.law.murdoch.edu.au/dtlj/1999/vol1_1/blakeney.pdf) 2.1 The ... which  
 enhanced its utility, was the **use** of HyperText Mark-up Language (HTML) ...  
[www.law.murdoch.edu.au/dtlj/1999/vol1\\_1/blakeney.pdf](http://www.law.murdoch.edu.au/dtlj/1999/vol1_1/blakeney.pdf) - [Similar pages](#)

[Digital Technology Law Journal](#)  
 ... was implied that Australia already had a **fair use** defence). ... Internet] URL: [http://www.law.murdoch.edu.au/dtlj/1999/vol1\\_2/bentley.html](http://www.law.murdoch.edu.au/dtlj/1999/vol1_2/bentley.html) - 29k - [Cached](#) - [Similar pages](#)

specification in Google cannot exceed the basic domain name. (Northern Light, which I also use for the same purposes, is more flexible in this regard, as you may specify any parts of the URL in any sequence.)

There is a good solution, however, in Google which works most of the time. Add a unique part of the preferred journal's URL to the topical query (fair use dtlj) in Google, and it will come back in a few seconds with hits just from *DTLJ*. Using both Northern Light and Google may be useful in cases where one covers larger backfiles and the other has more current coverage.

There is another big advantage in using Google for such searches. It searches and indexes also journals in PDF or RTF format. While for this journal it did not make a difference because the article is available both in HTML and PDF format, in the case of other journals, such as LASIE, the Australian library automation journal, PDF may be the only file format. The fact that all of my last

three examples are from Australian sources, is sheer coincidence.

In the long run all of the digital collections will have their native search engines, but until we reach that point indirect searching is a good solution. Even then, Google may be more efficient than some of the search engines that do not allow such features as exact phrase searching which is essential for full-text collections. On the other hand, Google, and other general search engines cannot help yet when the information is stored in a database format, and the result are produced in HTML format in response to a query. When visiting Web sites to harvest data, the spider components of search engines do not ask questions but just collect files they can recognise. Neither can they handle known file formats if access to them is password protected, so the solution is not universal, but still a useful tool in the savvy searcher's arsenal.

**Péter Jacsó**

*University of Hawaii at Manoa*

### Savvy searching columns, Vol. 25 Nos 4 and 6, 2001

The above issues featured discussion of databases, including Information Science Abstracts (published by Information Today Inc.)

EMERALD retracts any inaccuracies of fact about Information Science Abstracts, which may have been included in these columns.

The opinions expressed are those of the author, not necessarily the Editor or Publisher. EMERALD does not accept responsibility for the accuracy of information contained in the text.