How to Search and FIND – Deep Indexing of Figures and Tables

INFORUM 2007: 13th Conference on Professional Information Resources

Prague, May 22-24, 2007

Helle Lauridsen

Proquest CSA

UK

hlauridsen@csa.com

Abnstract

A scientific paper is a written report describing original research results whose format has been defined by centuries of developing tradition, editorial practice, scientific ethics and the interplay with printing and publishing services. The result of this process is that virtually every scientific paper has a title, abstract, introduction, materials and methods, results and discussion.

It is in this WRITTEN report the abstracting and indexing for A&I databases is normally done, not in the primary research results as represented by the tables and figures in the article.

This method has worked well for over a century as scientific information has been fairly limited, but with the information explosion[1] of the past 20 years information retrieval by traditional indexing was only temporarily saved by the speed and ease of electronic searching. At present we have so much information at our fingertips that the question is not to find it, but to limit it to the most relevant material.

Figures and tables represent the distilled essence of research communicated in academic articles. Although the analysis contained in the surrounding text is important, it is clear that researchers are eager to view the actual data collected, observed, or modelled to determine the article's relevance to their own work. The summary of data displayed in figures and tables is a highly valuable surrogate for the typically unavailable raw data sets..

The primary objective of a literature search is to locate information relevant to researchers' interests.

Neither traditional article-level indexing nor full text-level indexing where all text within a document is searchable can locate those publications which contain specific data of interest. By indexing the variables defined in tables and figures, researchers can find data with pinpoint accuracy.
For years this has been too difficult to do as many figures only appear as .jpg images and thus are closed for machine indexing. But two years ago the idea was revived by an innovative persons within the company CSA and now the technology had been developed to be able to do the actual indexing of the many figures and graphs in the articles.

The concept is that all tables and figures contained within an article are indexed. The number of records in a Tables & Figures Index is an order of magnitude greater than those contained in a typical abstracts database. In the database each record is being assigned one or more general categories reflecting the 'type' of data display (e.g. Photomicrograph, Histogram, Line Graph, Map of Study Site, and so on).
The figures are being Indexed – The primary terms enabling accurate searching:
a. Subject Indexing – Key variables presented in the figure or table
b. Geographic Indexing –A applicable geographic terms
c. Taxonomic Indexing – The Latin names of organisms will be included when appropriate, most

[1] In its first year of publication (1907) Chemical Abstracts contained a total of less than 12,000 abstracts. By contrast, Chemical Abstracts published a million abstracts just in 2006

often consisting of the genus and species names, but will include broader categories when available (e.g. family, class, etc.).

d. Statistical Indexing – Any standard statistical term relevant to a particular data display (e.g. Analysis of Covariance, ANCOVA, Simple Linear Regression, etc.)

e. Other Relevant Data – an indication of whether the table or figure contains either an empirical or theoretical predictive model

And finally each record is linked back to the source journal article

The perceived Benefits of Searching Tables &Figures was that:

-Targeted searches could be constructed by employing figure-oriented searches allowing the researcher to save time and match retrieval to specific data contained in the article.

-Researchers could ensure that the study actually focused on a specific variable, rather than simply referring to it indirectly (i.e. from another publication).

-Categories of objects could easily be browsed allowing easy creation of visuals for conference presentations, teaching or seminars.

But all this was theory and before launching this very large expensive project for real, CSA wanted to make sure that the idea was viable and in the spring of 2006 an in depth market research was initiated: to make it non-company related CSA asked a research team led by Professor Carol Tenopir, with Donald W. King, Dr.Robert Sandusky at the University of Tennessee, Center for Information Studies, to test the utility of deep indexing for scientists and explore how it might enhance scientific research

The team identified librarians at universities and research institutes in Europe and North America who would assist with the recruitment of scientists to test the system. In all, sixty scientists in 9 organizations participated (7 universities and 2 research institutes; 3 in Europe and 6 in the United States)

One member of the research team visited each of the participating organizations, to provide introductory sessions, gather data, distribute passwords, and provide instructions on additional data collection. Multiple methods of data collection allow data validation and triangulation for both quantitive and qualitive data. They allowed the team to study both predictive questions, such as how indexing of tables and figures might be used by scientists, and functional questions such as what type of search and interface features are particularly useful for a tables and figures system.

Data collection methods included: pre- and post-search questionnaires to describe potential usefulness, expectations and current practices; observation sessions to discover, through initial and real-time interactions with the system, potential usability and functionality issues; and structured diaries of searches performed by the participants, on topics of their own choosing in the weeks following the introductory sessions to gather more detail on potential uses of the Tables and Figures index prototype, encourage additional participant experiences with the system, and identify both useful functions and concerns with the prototype.

Electronic indexing / abstracting services were the most frequently used kind of resource: 35 participants (58%) indicated that 60% or more of their searches, and 49 (82%) reported that 40% or more of their searches were performed using electronic indexing / abstracting services. [2]

In other words, it was a highly experienced, information literate test panel, who had agreed to participate and this is some of the responses the survey team got:

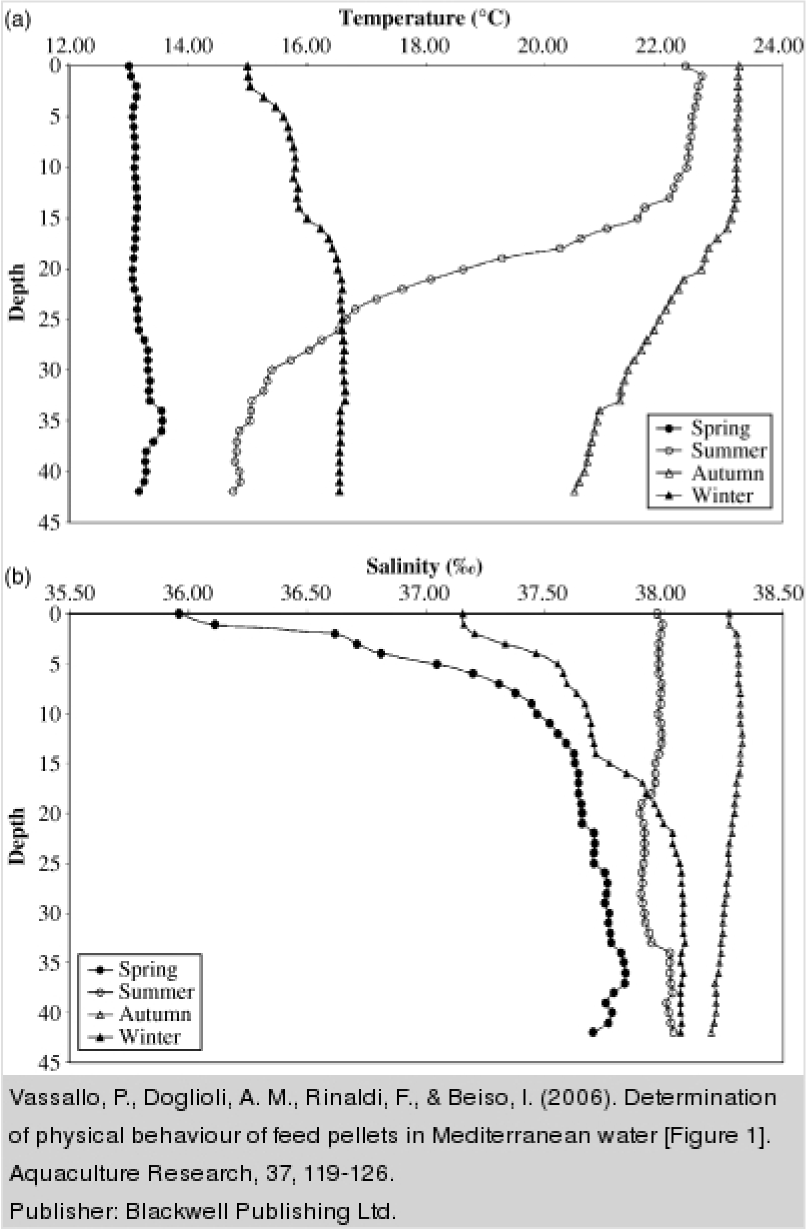
Overwhelmingly, participants alluded to the fact that this capability saved time and provided quicker access to information. “I can find the tables and figures that I need quickly, [and] it can save me a lot of time. I can work more efficiently” (Post Doc, Biology). One participant mentioned the increased efficiency of the search process, stating “It makes the search much quicker when it is focused” (Post Doc, Biology), and another noted that “the tables and figures are really helpful for scanning large sets of data first” (Post Doc, Oceanography). Some participants specifically noted that this quicker access and search time was a convenient aid to presentation preparation: “[i]t takes less time to find the information I want and especially I would find this useful when making a presentation” (Student, Biology). Another wrote: “I could find relevant information more quickly and images that were useful for presentations and research” (Professor, Engineering).[3]

But not all was favourable: Many participants commented on problems related to images and thumbnails. The prototype had some problems with the images not enlarging and some of the figures were too small and of too poor quality to be of any use. The quality of the captioning could also be improved and it was in general agreed that the entire caption was important and should be included.

During the fall of 2006 the entire prototype was taken to pieces: the images needed to become MUCH better, the captions clearer, response times to be kept at a minimum. Smaller thumbnails appeared in the search results and in January 2007 CSA Illustrata was launched.

Since then many researchers and librarians have been able to test the new indexing method and the feedback has been extremely overwhelming.
Firstly the already perceived benefits of being able to search in data not hitherto indexed and thus retrieving hidden information, is holding true, at a demonstration a researcher commented that it was impossible to find any information about the temperature in the Ligurian sea – a search was launched and a few seconds later a table giving exactly this but published in a paper on feed pellets in Mediterranean water was found.

[2] **White Paper written by Dr. Carol Tenopir, Dr. Bob Sandusky and Margaret M. Casado**
[3] ibid

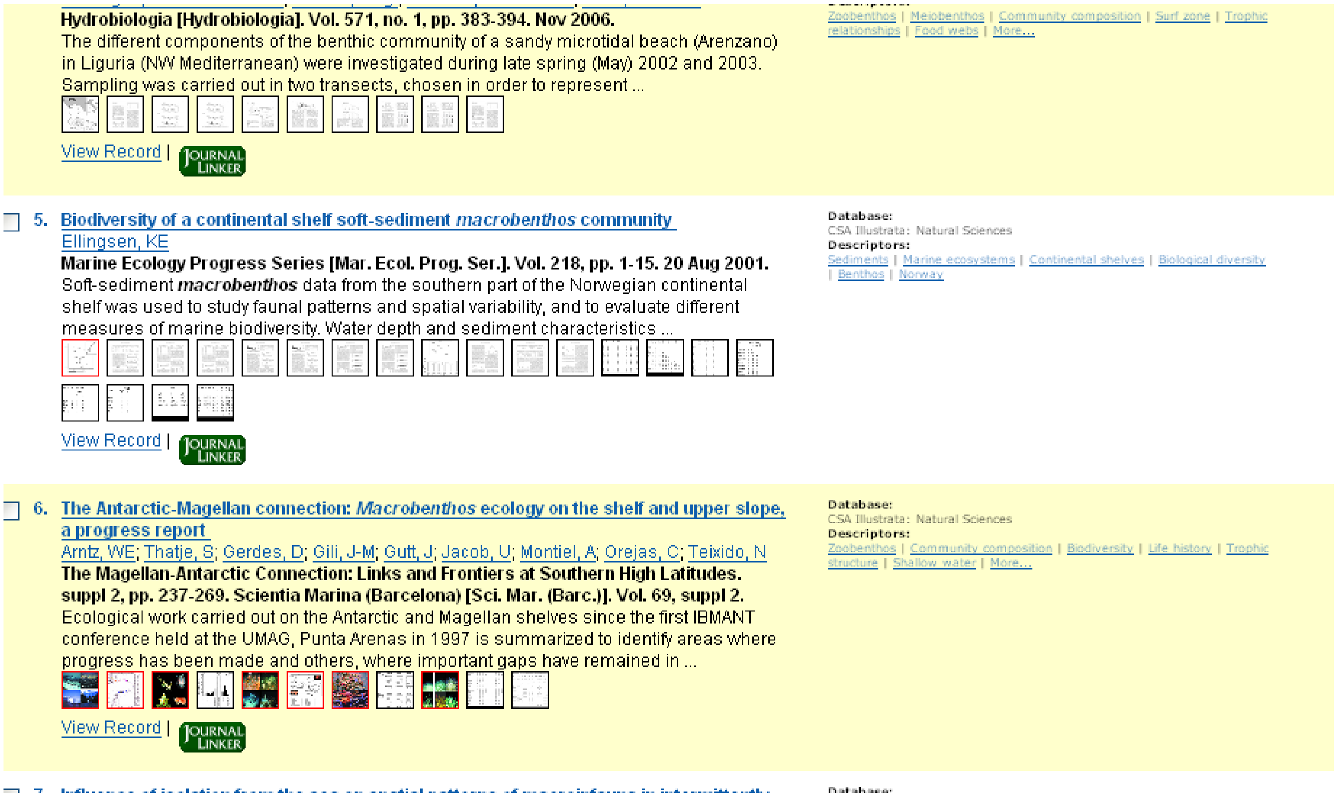


Vassallo, P., Doglioli, A. M., Rinaldi, F., & Beiso, I. (2006). Determination of physical behaviour of feed pellets in Mediterranean water [Figure 1]. Aquaculture Research, 37, 119-126.
Publisher: Blackwell Publishing Ltd.

Figure 1: indexing of information not formerly available

The rapidity with which you can find relevant illustrative information impresses most people seeing the database for the first time and the human brain and it's way of processing information should not be discounted in this matter: Only about 20% of the worlds population learns best from text based information[4], the rest from one of the other 4 learning styles – one of the most common of these being the image based.
Many people find it much easier to shift through search results including images and locate the relevant article fast and efficient

[4] Stephen Abrams, UKSG conference 2007



Hydrobiologia [Hydrobiologia]. Vol. 571, no. 1, pp. 383-394. Nov 2006.
The different components of the benthic community of a sandy microtidal beach (Arenzano) in Liguria (NW Mediterranean) were investigated during late spring (May) 2002 and 2003. Sampling was carried out in two transects, chosen in order to represent ...

View Record | JOURNAL LINKER

Descriptors:
Zoobenthos | Meiobenthos | Community composition | Surf zone | Trophic relationships | Food webs | More...

□ 5. Biodiversity of a continental shelf soft-sediment *macrobenthos* community
Ellingsen, KE
Marine Ecology Progress Series [Mar. Ecol. Prog. Ser.]. Vol. 218, pp. 1-15. 20 Aug 2001.
Soft-sediment ***macrobenthos*** data from the southern part of the Norwegian continental shelf was used to study faunal patterns and spatial variability, and to evaluate different measures of marine biodiversity. Water depth and sediment characteristics ...

View Record | JOURNAL LINKER

Database:
CSA Illustrata: Natural Sciences
Descriptors:
Sediments | Marine ecosystems | Continental shelves | Biological diversity | Benthos | Norway

□ 6. The Antarctic-Magellan connection: *Macrobenthos* ecology on the shelf and upper slope, a progress report
Arntz, WE; Thatje, S; Gerdes, D; Gili, J-M; Gutt, J; Jacob, U; Montiel, A; Orejas, C; Teixido, N
The Magellan-Antarctic Connection: Links and Frontiers at Southern High Latitudes. suppl 2, pp. 237-269. Scientia Marina (Barcelona) [Sci. Mar. (Barc.)]. Vol. 69, suppl 2.
Ecological work carried out on the Antarctic and Magellan shelves since the first IBMANT conference held at the UMAG, Punta Arenas in 1997 is summarized to identify areas where progress has been made and others, where important gaps have remained in ...

View Record | JOURNAL LINKER

Database:
CSA Illustrata: Natural Sciences
Descriptors:
Zoobenthos | Community composition | Biodiversity | Life history | Trophic structure | Shallow water | More...

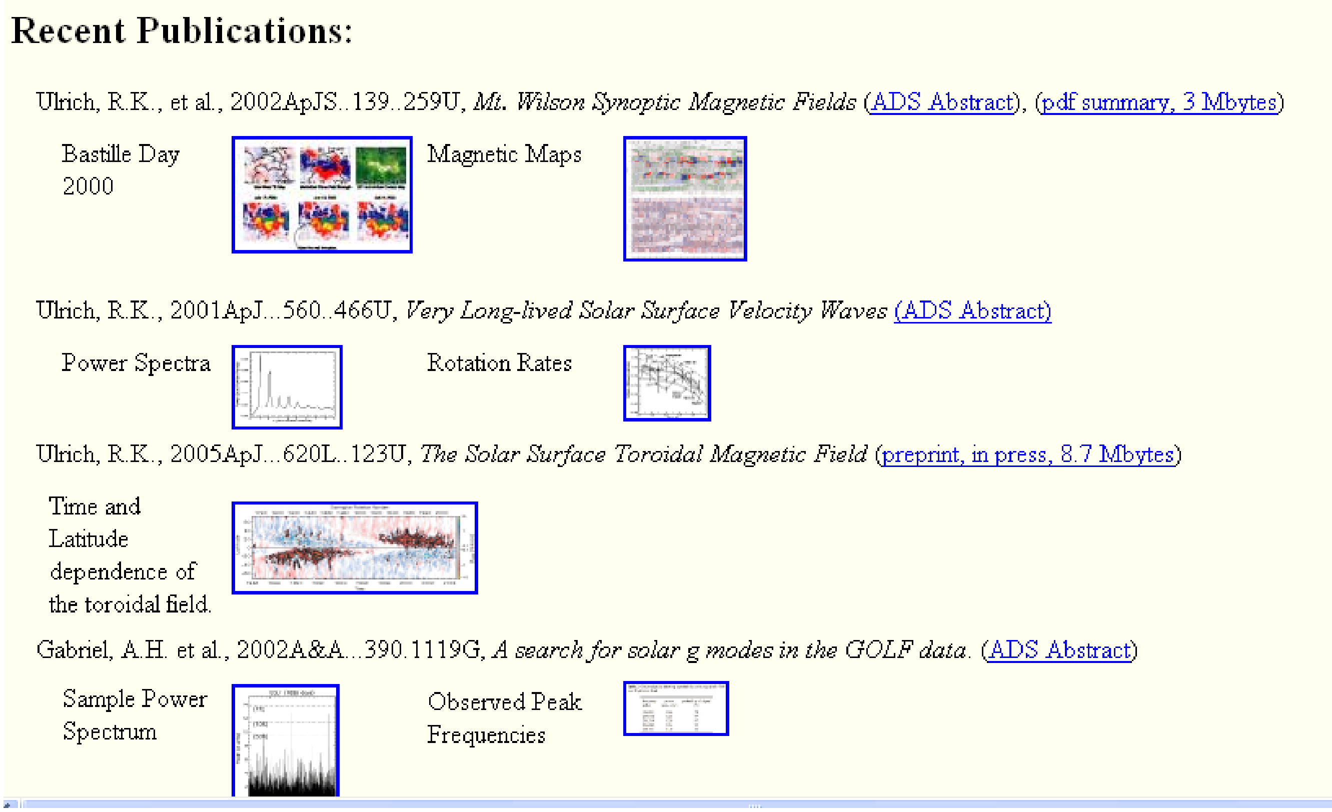
□ 7. Influence of isolation from the sea on spatial patterns of macroinfauna in intermittently

Database:

Figure 2: screenprint of search - easy to navigate by the human eye

Very often a database search is for an extension of already present knowledge and a quick view of the article images will easily indicate if the article has any interest.

This method giving a quick visual overview of the article content is also used by the researchers:



Recent Publications:

Ulrich, R.K., et al., 2002ApJS..139..259U, *Mt. Wilson Synoptic Magnetic Fields* (ADS Abstract), (pdf summary, 3 Mbytes)

Bastille Day 2000 Magnetic Maps

Ulrich, R.K., 2001ApJ...560..466U, *Very Long-lived Solar Surface Velocity Waves* (ADS Abstract)

Power Spectra Rotation Rates

Ulrich, R.K., 2005ApJ...620L..123U, *The Solar Surface Toroidal Magnetic Field* (preprint, in press, 8.7 Mbytes)

Time and Latitude dependence of the toroidal field.

Gabriel, A.H. et al., 2002A&A...390.1119G, *A search for solar g modes in the GOLF data.* (ADS Abstract)

Sample Power Spectrum Observed Peak Frequencies

Figure 3: Screen print of researcher's webpage

Conclusion:

A new way of deep mining information has been born. It is the hope of CSA that it will enhance the way literature is retrieved and help researchers and librarians alike to FIND relevant information rather than just retrieve huge search sets. Present days information explosion has long called for a different approach for retrieving information in a exact and timely way.