

**The value of structure in searching scientific literature**

July 30, 2004

Author:

G. Griffiths, Publishing Technology Manager, Scopus, Amsterdam, The Netherlands



## Abstract

The trajectory through the peer-review, publishing and library selection procedures gives “library” information a dimension of validity that is often too easily taken for granted. Information found in an academic library is there for a reason, and can be considered reliable, although at times it might be difficult to find. Traditionally, it has been the job of experts such as research librarians to find, evaluate and select the most reliable and relevant material for scientists.

This paper discusses some ways in which highly structured data can support searching and other tasks involved in finding relevant scientific literature.

## Introduction

In addition to learned books and journals, scientific abstracting and indexing (A&I) databases are an important source of information at libraries. Once these became available online, first on vendor CDs and remote access and later via Internet, an immensely powerful resource was created. Expert searchers could use the highly structured information to pinpoint exactly the information that was relevant to them. The disadvantage of these sources, however, is that they are expensive to create, may be isolated from other sources and can in some cases be difficult to use.

With the rise of the World Wide Web, interconnectivity and general use of new media, the number and variety of information sources has mushroomed, as have the means of accessing and disclosing them. There is a vast ocean of information available and it is not only librarians but also users who can now access everything from Google to specialized subject databases on platforms such as Ovid, Dialog, and EBSCO.

Furthermore, the library is no longer a place reached through a turnstile. It extends to our desktops and laptops; to the same browser we use to locate a restaurant or retrieve the latest weather report. This new degree of accessibility, together with the variety of possible sources, poses a new problem for library users, who must now learn to distinguish between different types, quality and sources of information.

Some important factors to bear in mind when searching for information:

1. *The validity of the information found:* The following succinct remark made by one medical librarian illustrates this: “...many people do not realize that not everything found on the Internet is true”<sup>1</sup>.
2. *The scope of the source selected and the search method applied:* The content selected for inclusion or indexing by any information source - be it an A&I database or a web search engine - is always a selection. No source contains “everything”. In addition, search engines are tuned to return results that are most relevant to their users. Hence, a mainstream web search engine is tuned to the needs of a large public; many of whom are interested in Harry Potter or Britney Spears (Fig.1). Similarly, dedicated specialty engines such as the science web search engine, Scirus, have more focused coverage due to the specific sources they select for indexing. Such specialty search engines are also tuned to

---

<sup>1</sup> Library Connect Newsletter, April 2004, Vol. 5 (1), p. 9. Librarians speak up: Hella Bluhm-Stieber.  
<http://www.elsevier.com/inca/publications/misc/672915libraryconnectvol5.pdf>

favour results appropriate to the aims of their target audience – in the case of Scirus, science searching.

Fig. 1: Top 5 most popular Google queries, 2003<sup>2</sup>.

Rank	Term
1	britney spears
2	harry potter
3	matrix
4	shakira
5	david beckham

3. *The purpose of the search (broad or specific)*: A student wanting to investigate a new, unknown topic will search more broadly and on different sources than a researcher wanting to make sure that he or she has not missed any published report of a newly observed effect. Sometimes a search task may be highly defined and the researcher or librarian will know exactly which source to check, will do so regularly, and will hence spot changes easily. But when orienting oneself in a new area, typically with a broad sweep of the web or a search on the main fields of a number of databases, the need quickly arises to narrow down one's search to more specific, relevant and trusted material. It is at this stage that it becomes useful to have a greater understanding of how to search.
4. *The structure of the data in a source*: The way information is structured strongly influences the results that can be obtained by any search method. For this reason, some idea of how the search works in conjunction with the data will help the user to improve and better interpret the value of the results.

### What is Structure in Data?

Information in a file or web page can be structured in a number of ways. At the simplest level, any text or document has its own inherent structure. This may simply be the native structure of the language in which it is written or the order of columns in a table. A particular type of document may conform to conventions that are reflected in it but not stated – for example a research paper starts with title, authors, affiliation, date, then perhaps an abstract, an introduction, procedures and results and then some discussion and a conclusion is followed by references and appendices of additional material.

This kind of implicit structure is useful in attempting to deduce the significance of parts of the text within the document. Search techniques involving linguistic processing and entity extraction make use of this, but the information so gathered is *deduced* and therefore less certain than if the information had been explicitly labeled as being the author or date.

On a more detailed level, a document may contain explicit physical structuring: for example, tags or labels splitting the document into named textual and logical elements. This indicates textual elements like title and body text, author or date, keywords, where an image should be placed, or metadata giving information about the document. The degree of structure can range from very simple to highly complex.

For example, an article may have a title, some body text, a creator and a date. This may be indicated by tagging in HTML or in a standard scheme such as Dublin Core<sup>3</sup>. The article may

<sup>2</sup> [http://www.clickz.com/stats/big\\_picture/traffic\\_patterns/article.php/5931\\_3293581](http://www.clickz.com/stats/big_picture/traffic_patterns/article.php/5931_3293581)

<sup>3</sup> Dublin Core Initiative - <http://dublincore.org/>

be marked up in complex XML, defining everything from title, authors and linked affiliations down to company names, compounds, and the last subscript in a mathematical formula.

An entirely different dimension of structure in data is an explicitly defined logical framework imposed by the environment in which the document is held or from which it originates. The framework may be the application of categorization, or a controlled index or thesaurus, which can be consulted or applied by the user or by software applications.

Such a framework relates not to the internal organization of a document but to the actual meaning of its content, and hence indicates the significance of the document within its environment and in relation to other documents and external concepts. As a simple example, Yahoo! assigns web pages to various categories and provides an option to narrow down a search by searching only within the category selected. In the most complex case, A&I databases like MEDLINE, EMBASE or Compendex employ experts to add meaning to records by indexing (typically using a thesaurus) and applying other categorizations or enhancements based on a common understanding of the content rather than on the raw text.

### **Different Search Techniques for Different Material**

The best way to find what we are looking for depends on the resource we are searching. In a highly structured A&I database, we know the boundaries of what is in there, i.e. how many items it contains and where they come from. We can also exploit the formatting and the relational structure in the database, both in our searches and as the basis of other functionalities. We can be accurate to the point of being sure that something is NOT in the database if the search does not find it.

If we search a source in which all items have the same structure, we can have great confidence that the outcome reflects our intentions. For example, if we search a bespoke database like PubMed for the author "May", we can be sure we will not get results where "May" appears in the title or abstract – or in the date. The information in this database is rigorously and richly encoded, and is the same across all records, so we know exactly what part of the record is the author, the abstract, the references, or the publication date, and can rely upon its consistency in searching.

To improve and understand our search results we need to have some knowledge of what the structure is – what a particular field or category includes/excludes and whether the structure has been applied consistently. To make optimal use of all of this detail and build precise and exhaustive searches, we need appropriate search strategies and tools. The classic example of traditional database searching is a finely tuned, watertight Boolean query developed by an information specialist, exploiting maximum detail on a single database or vendor platform. However, if we transfer such a detailed search to a different source in which the data is not structured according to the same rules, we will often get unexpected results. Although in science the most commonly used fields – like author, abstract, title and publication date – are frequently the same across many databases, the structure can vary immensely beyond this, particularly in the details of document structure and classification. This means that in a cross-database or federated search, the number of fields from any one source that can be used in the common search will be more or less restricted to those that are shared by all sources. Clearly, the more sources included, the smaller the shared set of searchable fields will become.

The extreme case of this is the web. A broad web search across billions of items in millions of formats can do no better than use the lowest common denominator. It is possible to restrict a search in the major web search engines on a number of metadata fields - such as

language, domain, date and file type - which are almost always indexed by the search engine. However, these are generally of little value in a scientific context. In many specialty search products, for example the science search engine, Scirus, it is possible to search within fields more closely related to article structure. If we search web data for “May” in the field “author name”, we are actually specifying that we want documents that have an explicit or implicit field for the author name. Furthermore, we also want this field to contain the word “May”, and hence we want to exclude anything that is not structured with a field “author” containing “May”. This can in fact be a useful strategy to find scientific articles only: employ the knowledge that most scientific articles have an author field.

It is important to remember that the vast majority of information available lacks useful structure. Structured and selected databases account for only a small percentage of all available data. So much research has gone into developing search methods to cope with heterogeneous and poorly structured data.

### **How Web Search Engines Cope with Lack of Structure**

The Web consists of a vast, shifting population of uncontrolled items with practically no shared structure that is useful in improving search results.

The predominant characteristic of web content is its diversity. The technology that has been developed to cope with this is quite specific to this kind of content. A vast amount of development is spent on “clever” web-oriented searching and information extraction, using advanced linguistic and statistical techniques on both the pages themselves and their interlinking with other pages. The web search software has to be “clever” because the material searched is so diverse and inherently unstructured, and the audience so wide. In this approach the aim is to retrieve sufficient relevant results for the user’s purpose rather than to be exhaustive.

Although there is evidence to show that multiple keywords are being used in web searches, there are still a high percentage of users using only one or two keywords<sup>4</sup>. Short queries of this type are very difficult to interpret. Query interpretation can be enhanced by prompting users to give more information or by inferring their intentions from previous search behavior. However, in general, the results from a simple keyword search will be numerous and varied, and so a major focus of the best web search engines lies in the proper determination of the relevance of any particular result to the original query.

Instead of just producing a list of everything found that contains the word or words specified, each hit is accompanied by a relevance score built from various components, depending on the data, the search algorithm used, and how it is configured or “tuned”. The results from the search are ranked according to a set of internal criteria so the searcher sees a list in which the items at the beginning of the list are *more likely* to fit what he or she (as average user) is looking for, than those at the end. In a very large result set this is vital: in a set of several thousand results, it is unusual for a searcher to browse past the second or third page<sup>5</sup>. Additionally, because the data lacks common structure, there is no other simple way to sort or handle results.

Any relevance algorithm may be composed of many different ranking factors. Relevance scores may be based on statistical and frequency measurements – for example number of

---

<sup>4</sup> NPD Search and Portal Site Study <http://searchenginewatch.com/sereport/article.php/2162791>

<sup>5</sup> “Impatient web searchers measure web sites’ appeal in seconds” <http://live.psu.edu/index.php?cmd=vs&story=3364>

occurrences of the search terms in the document in relation to the number of occurrences in the whole set; position of the terms in the document – early or late; whether the term appears in more crucial parts of the document, like the title; the number of links to or from the document from other sources and the relevance of the documents to or from which they link. Relevance scores can also be supplied by linguistically derived factors such as proximity of words to one another, deductions about document structure based on linguistic analysis, relative weights given to terms in certain parts of speech or in negative contexts.

Present and future development will continue to add to the repertoire of techniques and to the accuracy of searching large collections of data. Many enhancements of web searching are due to structure being added, either by richer crawling and indexing where possible – some data collections indexed by web engines may in fact be highly structured – or by using linguistic analysis to determine the hidden structure in text and identify useful elements.

However, it will never be possible for a web search engine to quickly return ALL and ONLY relevant information – and this is something that is vital to a scientific researcher.

### **How Abstracting and Indexing adds Value to Data**

Apart from being deeply structured (which is in itself a considerable investment and enhancement), the data in an abstracting and indexing (A&I) database has something special:

- 1) First, it has been selected – by humans – as being relevant and of sufficient quality for inclusion, both during the original peer review leading to primary publication, and for inclusion in the database itself.
- 2) Second, during the abstracting and indexing process, an intelligent and knowledgeable agent, usually a highly qualified human being (although progress is being made in developing software to do at least part of the work automatically) reads each article and assigns classifications and index terms according to understanding of the content, not simply based on the words that appear in the text.

Index terms or descriptors are controlled terms belonging to a known subject index or thesaurus. Any article may also contain uncontrolled keywords or title words provided by the authors, but the strength of controlled terminology is its uniformity across the entire database and, if part of a thesaurus, its meaning within the defined hierarchy. Depending on the database structure, the assigned terms may just be the preferred form of a number of variants in a simple normalized index, and as such will have synonyms and spelling variants.

In a more complex informational framework such as a thesaurus, the terms associated with a particular record occupy a position in the hierarchy. Any term may be sub-term of a broader term, and in itself possess narrower terms as well. In a complex thesaurus, terms may also be qualified in some way to express relations other than synonymy, like simple relatedness. Additionally, thesaurus terms in a record can be flagged to have a specific context that indicates, for example, whether the term has been included as representing the major focus of the research or to designate a link to related topics or materials used. These additions have the following effects on the data:

- Increased consistency within the data (use of controlled terms and descriptors, preferred forms, normalization)
- Indication of relatedness between different items and to external concepts (classification, subject areas, other documents)

- Indication of more complex relationships (placement in hierarchical index or thesaurus)

The resulting enriched data serves as a basis for highly superior searching and navigation of the available records.

### How Structure Enhances Search and Result Handling

In addition to the simple fact that it is easier to search reliably across content that is consistently structured in format and terminology, we can see the following ways in which structure enhances search and result handling:

- 1) *Exploiting a Hierarchical Index:* When searching and refining search results, ways in which a hierarchical index structure can be exploited to enhance searching and browsing include:
  - Mapping search words to preferred terms, so all word variants are covered in a single search;
  - Restricting the search to a limited subject area or a subsection of the hierarchy;
  - “Exploding” a search to include all narrower terms of the term used;
  - Broadening or restricting a search using related (including parent) terms and qualifiers;
  - Finding related documents in a controlled way (by e.g. searching for all documents that have the same set of controlled index terms).
- 2) *Exploiting Multiple Fields:* The richness of A&I data makes it possible to precisely narrow down a set of results to all and only relevant material, without ever having looked at the texts themselves. However, exploiting all this detail can be difficult to understand and to build into a search. First, finding the right articles can require complex searches, which are specific to the structure of the particular database or thesaurus being used. Second, there is the problem of using the correct syntax, which is usually different for each platform being accessed. In this way it can become a difficult task that only a highly experienced information specialist can do well.

Fortunately for the 21<sup>st</sup> century library user, there are other options. The fine detail found in A&I data can be exploited not only in the classical sense for expert querying, but can be combined with advanced search tools and techniques originally developed to deal with unstructured web data.

A search engine applied to highly structured data can create multiple indexes (one per field), each of which is highly detailed. This means that rapid search can be done on many different fields in the data, combining speed with depth and precision, and increasing the accuracy of relevance ranking. It also means that the results of a search can also contain this amount of detail, thereby allowing them to be manipulated, refined, or used with confidence as a basis for new searches or comparisons.

- 3) *New Functionality:* In a good search and navigation interface, the user can search or manipulate results in a way that is transparent and easily understood, leaving the complex querying or processing of results to the back-end system. In this way the user can benefit from both the detail in the data and the expertise of the search technology without needing expert knowledge of either.

Clustering and grouping techniques give an overview of what has been retrieved in the results in various categories of information. Again, linguistic analysis can be used to

extract salient terms from free text to categorize the content. However, explicit structure will be a more reliable basis for this clustering and grouping. For example, a list can be displayed of all the authors or journal titles found (and the number of occurrences) in the result set, giving the user an immediate overview of the scope of the result set. This can be taken further and used to refine the results by selecting authors or journal titles to include or exclude. If a database has a hierarchical index or thesaurus, this process can be extended even further and used to hone down a result set in fine detail and to organize and prune result sets.

- 4) *Beyond Original Query and Original Database*: A different approach might be to start with a narrow search and extend the results to include things not originally considered. By combining the structure in the data with sensitive relevance algorithms, we can search in a less specific way and in a truly serendipitous manner extend our horizons beyond what we expected to find. An example of this is the “more like this” functionality.

With the right approach, using relevancy geared to particular field types (e.g., using vector or “fingerprinting” techniques on text fields), we can develop insights and see new relationships within our field of interest and perhaps beyond.

Using an index or thesaurus improves precision and recall not only in its original environment but also, if applied with care, outside it. A good thesaurus expresses concepts and relationships that are relatively independent of the database to which it is applied, and might therefore be applicable to other, related contexts. For example, it can offer synonyms and show relations that a user might not otherwise conceive, and thus provide alternative or additional terms for a web search.

- 5) *Beyond Searching and Result handling*: Highly structured data has benefits beyond those of searching and result handling, including:
  - **Accurate linking**: If an item contains clearly labeled fields, the information from those fields can be used by a template to construct accurate links – for example an OpenURL link – to servers and web services. This can be used for a variety of purposes, for example to check if the full text of a particular database article is accessible via the library’s link resolving software, to request the full text through document delivery or to search for the full text in a different source.
  - **Analysis and insights**: With modern analytical software and data mining techniques, different items of data can be compared across various axes in large data collections. Relationships and correlations can be exposed that human beings could not immediately discern without statistical analysis; clearly, the better the structure of the information scanned, the more information can be gleaned from it. This can then be presented to the user in a useful format, such as in a report or as a visualisation.

## **Conclusion: Achieving the Best of both Worlds**

Scientists and students – the users of scientific, technical, and medical libraries – are interested in getting hold of information, and not in the methods used to find it. Furthermore, these users don’t want to put a lot of work into the process of finding what they need, and therefore a quick web search is second nature and often assumed to be sufficient.

In the world of research, however, a very important source of information is A&I databases. Although these databases represent a small percentage of the world’s information, A&I data contains a lot of added value: its content is usually high quality and enhanced by the expert

addition of controlled index terms and by the application of classifications and thesauri, and its high degree of structure permits exactness and specificity in searching, browsing, and related functionality.

So what should the researcher do after (or even instead of) a first broad orientation with Google or with Scirus, specialized in indexing scientific literature? Ideally, a library user should be able to easily move across the scale from a broad general web search to a narrow, appropriate, and specific A&I database search, without having to change platforms or syntax, or give up and seek help.

The solution lies in the development of search and browse tools and interfaces that apply the superior technology developed for web applications to highly structured data – enabling users to get the most out of large, high quality structured content collections without needing deep knowledge of search techniques, syntax, or data structure. In this way, accurate searching will be made easier, while state-of-the-art analysis, linking and data manipulation techniques can be applied and improved due to the rich structure of the data.

### **Acknowledgements**

IJsbrand Jan Aalbersberg, Harriet Bell, Ian Crowlesmith, Ginny Hendricks, and Craig Scott for their contributions to this paper.