

## **DIVING FOR PEARLS: CONTROLLED SEARCHING ON EdNA ONLINE**

**Albert Ip**

Technical Specialist,  
EdNA HE Team, Australia.  
[albert@dls.au.com](mailto:albert@dls.au.com)

**Prof. Iain Morrison**

Dept of Information Systems,  
The University of Melbourne, Australia  
[i.morrison@dis.unimelb.edu.au](mailto:i.morrison@dis.unimelb.edu.au)

**Michael Currie**

EdNA Higher Education Project Manager  
EdNA HE Team, Australia.  
[m.currie@dis.unimelb.edu.au](mailto:m.currie@dis.unimelb.edu.au)

**Jon Mason**

Technical Director,  
Education.Au Limited, Australia  
[jmason@educationau.edu.au](mailto:jmason@educationau.edu.au)

### **Abstract**

*The recent release of a metadata toolset by the EdNA Higher Education Project team has provided an opportunity for the Australian education community to improve access to a wealth of information resources through EdNA Online and related services.*

*This paper is concerned with strategies for resource discovery. An evaluation of current Web-based search services reveals a need for greater precision, coverage, consistency and quality control. Subject Gateways, which typically utilise metadata to enhance their service, provide a potential solution and these are reviewed in the context of the principal author's data model.*

*EdNA Online demonstrates these qualities and this paper surveys the potential impact on its search capability of the new toolset and accompanying thesaurus. The authors conclude by assessing a number of possible developments, including metadata mapping and user profiling, that would serve to enhance resource discovery via EdNA Online.*

### **Keywords**

*resource discovery; subject gateways; metadata; online searching; World Wide Web; education; EdNA Online; user profiles; search engines*

### **Introduction**

The dramatic change in our capacity to encode, transmit and locate information in a digital form has triggered a paradigm shift in teaching and learning towards enquiry-based and resource-based learning techniques leading towards information-dense learning environments. The availability of rich sources of information will have significant impact on the design of these learning environments (Ip and Naidu, 2000). Not only will these information resources be central to teaching and learning in each level and subject area, the skills to discover, acquire, interpret, manage and use information will be key competencies in a learning society.

Providing access to high quality online resources is one of the main organising principles of EdNA Online, the online service developed by EdNA (Education Network Australia), a collaboration involving all education and training authorities throughout the Commonwealth, States and Territories of Australia. EdNA's primary focus is to promote the effective use of the Internet in education and training. EdNA Online is positioned as a comprehensive portal to information resources relevant to Australian education.

### **Evaluation of Search Engines**

The small amount of literature evaluating the information retrieval process on the Web has related primarily to commercial search engines. These mainly use an inverted index – a keyword histogramming technique with relevance measured to the number of keyword 'hits' registered over the search query, as a means of quickly locating and retrieving relevant information.

While commercial search engines are most commonly used for accessing Web-based information, various studies (Chu and Rosenthal, 1996; Lawrence and Giles, 1999) have revealed that these services are neither comprehensive nor accurate. Search engines can be assessed in terms of coverage, recall, consistency and precision.

*Coverage* is about how comprehensive search services are. Studies by Lawrence and Giles (1999) and Broder *et al* (2000) indicate that the number of individual pages on the Web exceeds 1.6 billion. Current estimates reveal that the largest search engine, *FAST*, includes some 300 million pages while *Alta Vista* and *Excite* list around 250 million each. (Sherman, 2000). The largest search engines today therefore cover only about 16% of estimated Web pages available.

But concerns about adequate coverage are not limited to raw size. Broder *et al* (2000), using *Alta Vista* crawl data covering 200 million pages, described the whole Web as a "bow" structure with a strongly connected core component (SCC) at the centre and an *IN* and an *OUT* section as the "bows". The *IN* section represents pages that have links to the core, but no links back to them. The *OUT* section represents pages that are linked from the core but do not have links back to it. The SCC and *OUT* regions together were about 99 million pages out of the 200 million test pages. These were roughly the pages directly indexable from the home pages.

The analysis by Broder *et al* highlights the problem of "page islands" - pages that are unreachable by the uni-directional linkage of the Web. Unless these "page islands" are submitted directly to search engines, there is no way "web crawling" will discover them. This suggests that less than 50% of indexable documents are likely to be reached by search engines using robots to access home pages.

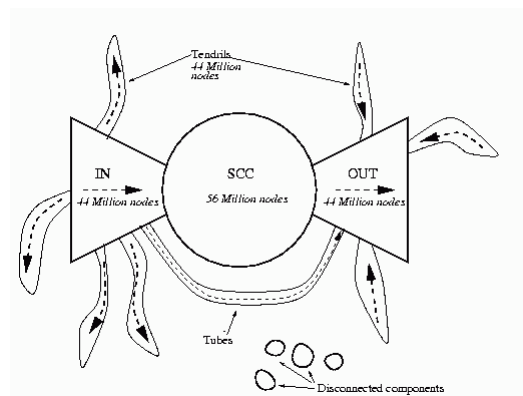


Figure 1: Web structure (Border *et al*)

The discovery of pages contained within the *IN* component reflects the action of information providers actively promoting their resources. What about those resources the original creators do not actively promote which are located in the page islands? On this basis, no search service can state that it has significant *coverage* of resources available on the Web.

*Recall* involves the ability of search services to return results so as to meet the needs of the searcher. Typical results from search engines can often return thousands of resources even with the current level of *coverage*. Few searchers will work through all the returned resources. This introduces the issue of ranking systems used by different engines. While a wide range of strategies are used (Ip *et al*, 1999), many of these techniques are at best a superficial effort and increasingly prone to error as measured by the failure in relevance of returned resources (Curtois and Berry, 1999.).

One reason for this is the commercial value attached to having a high ranking on major search engine lists and the increasing incidence of spamming techniques that are designed to artificially enhance ranking positions in search indexes. This includes the use of extraneous, irrelevant and duplicated words purely designed to confuse ranking algorithms (Notess, May 2000).

*Consistency* is another area that has suffered in the desire of general search engines to be the largest and fastest. Consistency measures the repeatability and coherence of search query results over short timeframes. As Notess has stated, "Unfortunately for the consistency connoisseur, many of the less well-known Internet search tools are hastily constructed. Meanwhile, the top Internet search engines have had extensive development of their interface, but with the general aim of providing a few relevant answers quickly to almost any kind of search. In neither case is search consistency necessarily a high priority." (Notess, March 2000)

*Precision* measures the efficacy of the search – that is, the capacity of the search process to find resources that clearly match the search criteria. Leighton and Srivastava (1997) did precision studies on index services of five commercial search engines during general subject queries in the academic setting. Results indicated that if the "criterion is a page that is very likely to be useful, the median disappears down to 0.06" in the possible range of 0 to 1. The researchers ascribed this to the lack of features enabling clarification of search terms.

### **A Data Model for Online Information**

Ip *et al* (1999) developed a data model to clarify the relationships between the different types of information used by search engines. Primary information is referred to as Type-1 data. This is the data contained in the resources information seekers are interested in. However, most search engines normally do not store the original resource.

Instead, search sites create or use Type-2 data to support their operation. Commercial engines are mainly based on an inverted index of the resources whereas metadata search sites such as EdNA Online are based on metadata describing the resources. Ip *et al* define the organisation of, and the relationships between groups of Type-2 data (especially in the case of metadata) as Type-3 data.

### **Subject Gateways as Resource Collections**

The sustainability of specialized search sites, such as EdNA Online or subject gateways deserves special consideration. Such sites are normally characterised by relatively small collections and associated resource descriptors, the use of sophisticated metadata and the adoption of metadata standards. They offer a high degree of selectivity towards resources that offer special value to their identified community.

It is becoming obvious that it is difficult to really trust or expect much refinement or focus from commercial search engines given the ever-increasing size of the search space and the existence of the significant *IN* component (Broder *et al*, 2000). While many depend on the "recommend URL" services, this strategy for reaching the "page islands" will be buried in the chaos of other information.

The value of any subject gateway depends upon efficient and effective ways to build, index and facilitate retrieval from special interest collections sifted or culled from the massive underlying search space. This sifting and culling depends on the domain knowledge of the subject gateway owners. Such domain expertise is hard to find in general purpose search engines. More specialized search sites, such as EdNA Online, tend to place a higher value on holding and returning quality resources than general search engines. To achieve and maintain this quality, specialized search sites typically depend on metadata, the coverage of the resources within the confines of their charter, the quality of the Type-1 resources being included, and the ability to return results appropriate to the needs of their clientele.

In their research, Lawrence and Giles (1999) found that 34.2% of Web servers contain some pages with metadata tags in the HTML source. There were no further data on the proportion of pages on such servers that have metadata. Standardised schemes, such as Dublin Core were only used in 0.3% of sites. The overall uptake of metadata by page counts would be a much smaller number than that quoted by Lawrence and Giles (1999).

It is no wonder, then, that major commercial search engines which compete according to the size of their Type-2 data collections are based on an "inverted index" as their Type-2 data creation method. Creating inverted index Type-2 data is an automatic process requiring no human intervention. There is very little user-provided indexing assistance in the form of metadata elements and keywords/phrases that help identify the semantics of the resource and intended audience/use.

Most subject gateways utilise sophisticated metadata schemas as opposed to a general dearth of metadata on the Web, as noted by Lawrence and Giles (1999). Ironically, the manual creation of metadata – especially detached metadata – is an important mechanism for subject gateways to maintain quality and value to their community.

The major issue faced by search engines in their Type-2 data representation of the collection is the lack of a semantic context to how a keyword or phrase is used in the original Type-1 resource. This will have significant impact on the ability of search engines to rank the search results. Metadata potentially can solve this issue by developing a sophisticated ontology of the classification scheme and mapping keywords into concepts utilising a thesaurus or 'open' or 'controlled' vocabulary.

### **Information Retrieval on EdNA Online**

The resources available via EdNA Online can be currently retrieved in any of three modes: Searching, Browsing and more recently, Pathways. Browsing and Searching are implemented conventionally while Pathways is based upon a more advanced concept that works by generating a dynamic set of possible paths after each step that the user takes in their search.

The *browse* category structure is a hierarchical guide into the EdNA Online information space. Thus, the category tree starts with a sector (education sector – school education, VET or university) view. The Pathways project is aimed at providing alternate paths into the information space based on the user's immediate past or historical preferences or navigation history, hence the name 'Pathway'. The idea is to dynamically generate the next level of branches as users are navigating into the information space.

Search functions on EdNA Online are built around two collections of Type-2 data. The core collection is built from metadata records evaluated or developed by reviewers). A secondary (inverted) free text index is created as a result of "spidering" the core collection. Searches can be specified based on keywords against the metadata record or against the inverted index.

The increase of the search space of EdNA Online depends on both the 'core' and 'non-core' items. There are three mechanisms of increasing the 'core' items collection:

- submissions from institutions which have service agreements with *education.au limited* (the managers of EdNA Online),
- suggestions from the online users, and
- resources manually discovered and evaluated by Directory Officers.

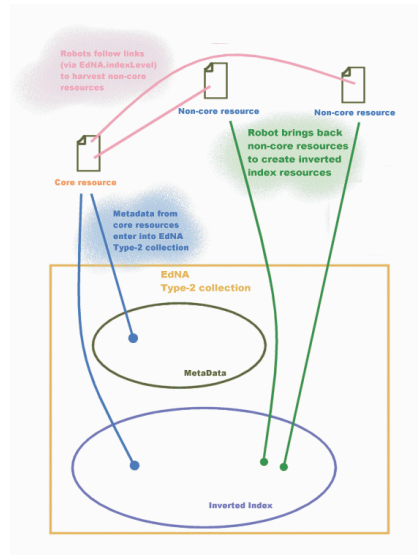


Figure2: Relationship between core and non-core items on EdNA Online

### Improving Search

While commercial search engines are becoming increasingly refined, most still use an inverted index for searching. Simple keyword searching of these search engines typically return thousands of pages covering widely different semantics and contexts, e.g., the word "conductor" could return references to persons who collect fares as well as references to metal that conducts electricity. This is not a very efficient search technique.

Feldman (2000) advocates the use of other types of technologies to meet the information retrieval needs including natural language processing, intelligent agents, concept extraction and mapping, machine-aided indexing, entity extraction, relationship extraction and text mining. These are emerging technologies and it is important for EdNA (the network) to be informed by progress in these areas from both 'use' and 'technology' perspectives. For pragmatic reasons, many efforts that are currently in progress advocate the use of metadata to support resource discovery on the basis that the originator of the resource would have had knowledge about semantics and intended audience/use.

Metadata has been used to describe, organize and support retrieval of information. Ip *et al* (2000) advocate the use of metadata and linkages using metadata to support the "education-enabling" of resources which are not originally created for educational purposes.

### Resource Categorisation

Until now metadata used by EdNA Online has been collected from stakeholder collections and from international sources manually by directory officers and through local contributors. As EdNA Online continues to develop, it is reasonable to expect that the sheer number of resources will necessitate alternative strategies.

Based on the model suggested by Ip, *et al* (1999), the quality and coverage of any metadata-based search is only as good as the quality and breadth of the Type-2 data used to support it. The size of the collection (i.e., breadth) of Type-2 data determines the

coverage or the comprehensiveness of search results. The quality of the search results (i.e., appropriateness and precision) depends on the "quality" of Type-2 data supported by the ranking algorithm in use by the search engine.

One of possible "quality" improvement is the accurate classification of resources according to authoritative (well-structured) schema such as the Australian Standard Classification of Educational Descriptors (ASCED) or the Library of Congress Subject Headings (LCSH). The EdNA Metadata Standard, based upon Dublin Core, now accommodates such classifications.

### **Resource Descriptor Maintenance**

In 1999, the EdNA Higher Education project team was funded by DETYA to produce a suite of metadata tools for educational sectors and institutions in Australia. These tools were aimed at simplifying and partially automating the creation and maintenance of metadata for web resources initially intended to be added to EdNA Online.

The EdNA Higher Education project team also aimed to improve search effectiveness through the use of an educational thesaurus. This would be used by indexers to produce controlled vocabulary values within the *DC.Subject* element from keywords automatically selected by the metadata toolset from the resource. It was expected that the use of an EdNA specific thesaurus could enable the toolset to suggest appropriate values for the metadata element *EDNA.Categories*, which would also greatly improve the accuracy and efficiency in assigning values to this element. A test EdNA thesaurus was developed based on a review commissioned by *education.au limited* in 1998-9. The tools support changes to or replacement of the thesaurus.

As part of the improvement to the information retrieval service, it is also anticipated that the thesaurus would make possible better query matching.

### **Using a Thesaurus with EdNA Online**

Although server-side support of the use of a thesaurus for EdNA Online has not yet been implemented, the thesaurus can be applied at the user end. The thesaurus viewer from the EdNA metadata tool set can be used in a stand-alone manner. By using the viewer it is possible to find the keywords that are actually being used. We expect there would be significant improvement of search results, especially in relationship to precision, when the historical Type-2 data currently used by EdNA Online has been updated using the new tool.

### **Future work**

The reviewed evaluation studies on commercial search engines did not look into the underlying operating support data (i.e., Type-2 data) and ranking algorithms. These studies focussed on the overall usability of the search engines. As a specific study to find out the effect on the use of thesaurus on user experience, it is important to establish the improvements to the quality of the Type-2 data, and the search results.

It has been established that most users only use the beginning portion of search results. The overall user experience is also dependent on the ranking algorithm, the method by which results are presented in order of relevance. By expanding the scope of the study to include the ranking algorithm, the report will be of much higher value.

In the meantime, a more comprehensive literature study needs to be performed to determine acceptable metrics for determining the usability and experience of information retrieval for systems based on metadata. It would be informative if such metrics can also provide a comparison of EdNA Online with commercial inverted index-based search engines.

After a significant number of resources on EdNA Online have been created using the thesaurus, the measurement can be performed again to establish a post-application result. It can be compared with the established benchmark to determine the effectiveness of the metadata tools.

### ***Type-2 Data Quality Metrics***

The effectiveness of a search engine, irrespective of the underlying mechanism, is to be judged by the final user experience. Among the criteria proposed by Lancaster and Fayen (1973) "Coverage" and "Precision" are directly related to the underlying Type-2 data quality. Leighton and Srivastava (1997) used a scale rating the "relevancy" of the links and their formula included a measure of the number of such links weighted by their relevancy scale.

While "coverage", or the size of the Type-2 data hosted by EdNA Online is a matter to consider on another occasion, the "precision" or "relevancy" is highly dependent on whether the appropriate values have been assigned to the appropriate metadata elements.

The choice of a thesaurus based on the EdNA Online category structure for the EdNA metadata tools provides the opportunity to automatically enable and support the creation of values for these elements. A metric to establish the effectiveness is needed.

### ***EdNA User Profiling***

Subject Gateways obtain their value from their ability to meet the needs of their stakeholders. The concept of User Profiling, then, has strong appeal as it applies to the interface design and to the gateway's ability to provide a customised service, targeting the previously determined needs and interests of the user. Already the EdNA Pathways project has gone some way to achieving this. A user can run a search and then save the pathway to be used later for a search on the same topic, either by the user or as a class activity.

As a cross sectoral service, the issue of providing search options and interface features specific to sectors and particular groups is being discussed and will become a major focus in the near future.

### ***Summary***

In an expanding information economy, students need to be able to efficiently access pertinent information. New information-dense learning environments will also be created. Subject Gateways, with their focussed and accessible collections, can provide a valuable resource.

However, the quality of any search tool (including comprehensiveness, precision and ranking) depends on the underlying Type-2 data that supports the search or browse functionality. As part of the process of improving the quality of information retrieval, the

EdNA Higher Education project team believes that using an appropriate EdNA-specific thesaurus is a significant step in the right direction.

The team has now completed the delivery of a toolset which enable the use of a thesaurus during the creation process of Type-2 data. Final effectiveness or improvements will depend on the wide acceptance of the toolset, the updating of the historical Type-2 data currently used by EdNA Online and implementation of a matching thesaurus on the search side of EdNA Online.

This paper proposes a possible step forward. We outlined an evaluation framework to benchmark potential improvements to the search process. Ultimately however, the effectiveness of search on EdNA Online must be measured against its ability to meet the specific needs of a wide cross-section of educators and students.

## References

- Broder,A., Kumar,R., Maghoul,F., Raghavan,P., Rajagopalan,S., Stata,R., Tomkins,A. and Wiener,J (2000) Graph Structure in the Web. URL: <http://www.almaden.ibm.com/cs/k53/www9.final/>
- Choo, C.H., Detlor, B., and Turnbull, D. (Feb. 2000) 'Information Seeking on the Web: An Integrated Model of Browsing and Searching', First Monday, 5:2. URL: [http://firstmonday.org/issues/issue5\\_2/choo/index.html](http://firstmonday.org/issues/issue5_2/choo/index.html)
- Chu, H., and Rosenthal, M. (1996) 'Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology', ASIS 1996 Annual Conference Proceedings. URL: <http://www.asis.org/annual-96/ElectronicProceedings/chu.html>
- Courtois, M. and Berry, M.(May 1999). 'Results Ranking in Web Search Engines', Online, 23:3. URL: <http://www.onlineinc.com/onlinemag/OL1999/courtois5.html>
- Currie, M., Moss, N., and Ip A. (2000) 'The EdNA Metadata Toolset: a Case Study'. AusWeb2K The Sixth Australian World Wide Web Conference, Cairns URL: <http://ausweb.scu.edu.au/aw2k/papers/currie/index.html>
- Ellis, D. and Haugan, M. (1997). 'Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment'. Journal of Documentation, 53:4, p. 384-403.
- Feldman, S. (May 1999). 'NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval', Online, 23:3. URL: <http://www.onlineinc.com/onlinemag/OL1999/feldman5.html>
- Feldman, S.(Jan 2000), 'The Answer Machine', Searcher, 8:1. URL: <http://www.infotoday.com/searcher/jan00/feldman.htm>
- Ip, A., Currie, M., Morrison, I. and Mason J. (1999) 'Metasearching or Megasearching: Toward a Data Model for Distributed Resource Discovery' in Castro, F. et al. e-Education: Challenges and Opportunities: Proceedings of the Fifth Hong Kong Web Symposium. p. 65-82. URL: <http://www.dls.au.com/metadata/DataModel.html>

Ip,A., Morrison,I., Currie,M. and Mason,J. (2000) 'Managing Online Resources for Teaching and Learning' in Treloar,A., and Ellis,A., (ed.) The Web: Communication & Information Access for a New Millennium, (Proceedings of AusWeb2K, the Six Australian World Wide Web Conference, p 157-166

Ip, A.,and Naidu, S. (2000) 'Reuse of Web-Based Resources in Technology-Enhanced Student-Centered Learning Environments', Full paper submitted to IFET journal.

Lancaster, F.W., and Fayen, E.G. (1973). Information Retrieval On-Line, Los Angeles, CA: Melville Publishing Co., Chapter 6.

Lawrence, S. and Giles, C.L. (July 1999). 'Accessibility of Information on the Web', Nature, Vol. 400, p.107-109.

Leighton, H.V., and Srivastava, J.(1997). 'Precision among World Wide Web Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos'.URL: [www.winona.msus.edu/library/webind2/webind2.htm](http://www.winona.msus.edu/library/webind2/webind2.htm)

Notess, G. (March 2000) 'Search Engine Inconsistencies', Online, 24:2 URL: <http://www.onlineinc.com/onlinemag/OL2000/net3.html>

Notess, G. (May 2000). 'Up and Coming Search Technologies', Online, 24:3 URL: <http://www.onlineinc.com/onlinemag/OL2000/net5.html>

Sherman, C. (May 2000). 'The Future Revisited: What's New with Web Search', Online, 24:3. URL: <http://www.onlineinc.com/onlinemag/OL2000/sherman5.html>

## Copyright

Copyright © 2000 Albert Ip, Iain Morrison, Mike Currie and Jon Mason.

The author(s) assign to ASCILITE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The author(s) also grant a non-exclusive licence to ASCILITE to publish this document in full on the World Wide Web (prime sites and mirrors) and in printed form within the ASCILITE 2000 conference proceedings. Any other usage is prohibited without the express permission of the author(s).