Richard Lee Altman. Searching the IRN: Metadata at a Corporate Digital Library. A Master's paper for the M.S. in I.S. degree. April, 1999. 62 pages. Advisor: Diane Sonnenwald.

This study describes the implementation and evaluation of a search tool for online market research at a large corporate library. The author details the creation of a search system using metadata to catalog online documents, based the following open standards of document description: RDF, XML, and Dublin Core. This search system was implemented in a corporate library called the Information Resource Network (IRN). The effectiveness of this search tool was evaluated using an online survey, in combination with search log analysis. The survey was conducted to determine the search preferences of the IRN staff, and to measure their level of satisfaction with the new search engine.

Distribution of the survey to the IRN yielded a 67% response rate from the department's 38 professional researchers. Delivery of the surveys via a web-based corporate tool allowed extremely rapid response time, with the majority of participants responding within two days of notification. Three distinct user groups were identified within the IRN, based on job function: information specialists, associate specialists, and administrators. Analysis of the survey data and log files indicated that the new IRN search engine was well received. During the first four months of usage, 56.1% (23715) searches involved some combination of metadata options. Users voiced a common desire for wider coverage of online materials, with better notification of exactly which documents were searchable. Future search developments should focus on the creation of a universal metadata scheme for all online documents, expanding on the current version of the IRN search engine.

Headings:

Corporate Libraries

Special Libraries

Information retrieval

Information retrieval -- Case studies

Metadata

SEARCHING THE IRN:
METADATA AT A CORPORATE DIGITAL LIBRARY

By
Richard Lee Altman

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April, 1999

Approved by:

____
Advisor

Table of Contents

List of Figures

I. INTRODUCTION

The Problem

Along with the explosive growth of networked information, as exemplified by the Internet, has come great confusion about how to find the *right* information. Every Internet user knows the frustration of using all the major search engines, never knowing if the best documents have been found. Such engines as AltaVista demonstrate the limits of traditional Boolean-style searching, while more "concept-based" alternatives have shown no better performance. One academic study of major Internet search tools revealed that " even the best search services deliver only an estimated median of 10% … precision for very useful pages." (Leighton & Srivastava, 1997). None of these options provides anything resembling a guarantee of completeness for search results. With internal corporate networks of thousands of Web servers and millions of Web pages, many corporations face the same problems. For these corporations this is more than just an inconvenience; the productivity and profitability of the company can suffer from faulty distribution of information. In an international company where brainpower is the key resource, the efficient sharing of information is a top priority.

The Question:

In this study, I worked with a major corporate research library to investigate the best way of searching their corporate online resources. I focused on methods that would go beyond the limitations of simple full-text searching. Could the corporate research library offer a web-based search interface capable of satisfying professional reference specialists? Could I combine the strengths of library catalogs with the strengths of Internet search tools, so that employees worldwide could access the corporate electronic documents?

A Possible Solution

Metadata offers a possible solution to this problem by providing a framework to catalog digital resources. Though most of us have never heard of metadata, we have used it all our lives. Every time we use a card catalog, we utilize metadata. A "subject" card is value-added data that *describes* the source we seek (this source is itself a type of data). Loosely translated as "data about data," metadata provides a tool for describing online information resources.

Currently there are several efforts underway to create international "standards" for metadata, which would allow universal interoperability of metadata-based searching. Metadata removes the burden of retrieval accuracy from the search engines, and places it on the shoulders of site designers and cataloguers. If web designers and cataloguers add these tags, pages become self-describing. With metatags available, search engines need not guess at the content of a document. Metadata promises to democratize the search process, by allowing the creators of documents to add their own metatags for search engines to discover and index. For practical reasons such as the lack of a universal controlled vocabulary, it seems unlikely that metadata will be adopted rapidly by the general public or average person who builds a web page. However, in the hands of trained librarians, metadata can serve as a foundation for digital libraries of the future.

An Approach to a Solution

At a major international corporation, I created a metadata classification and search system for a digital library of market research. This corporation allocates a multimillion dollar budget for these research reports about the current state of the market, and it is crucial that employees have immediate access to these documents. Working with a professional search team at the corporation, I determined the best format for this metadata, and I created a workable metadata-based search system. The following operational questions guided my work:

- What information should be stored in metadata tags for this corporate archive?
- How do these metadata fields differ from traditional cataloging information?
- Did the addition of metadata improve search efficiency for trained information specialists?

These questions overlap to deal with different facets of a wider issue: What is the best way of implementing new standards of metadata for improving access to information in a digital library?

II. BACKGROUND AND LITERATURE REVIEW

This project aimed at delivering practical results for clients of a "virtual library" known as the Information Resource Network (IRN). The IRN web site serves as a central point for information distribution within the corporation, accessible by more than 70,000 employees worldwide. These end users vary widely in their technical skill level, so that web site design must reflect this variety of incoming knowledge. Because the search engine is the single universal access point for more than 30,000 online documents, I focused on creating and evaluating a search system that is intuitive, flexible, and powerful.

The IRN staff, consisting of some 75 full-time employees in seven international locations, is the single most influential group of stakeholders in the IRN search engine. 38 of these employees are research specialists who answer daily information requests from corporate managers, engineers, and marketers. Until December 1998, these specialists could access the IRN's online materials with two methods:

1. Full-text searches with the IRN's commercial search engine.
2. Traditional OPAC searches with the IRN's Infocat catalog.

Specialists did not trust either of these methods, for different reasons. The full-text search engine was implemented "out-of-the-box" with no customization, and tended to produce incomprehensible result lists. “Relevance” was the only available sorting option, and the ranking numbers made little sense to users. Apparently these relevance numbers were not based on frequency or proximity of query terms; rather, documents with high numbers of *synonyms* for query terms received the highest ranking. No date information was available. IRN specialists complained frequently that this search engine was unusable, because of its inability to produce meaningful result lists. Meanwhile, the Infocat OPAC offered no access to full-text documents, and depended on cataloguers for manual creation of entries. The cataloguing team could not keep pace with the arrival of new online materials, so that catalog searches seldom produced the latest documents. IRN research specialists needed a new method of searching, which could allow them to unify their search strategies with a single tool.

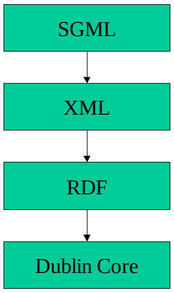


Figure 1: Hierarchy of Metadata-Related Standards

Metadata offered a technique for achieving this goal, by combining catalog data with full-text documents. A number of articles have addressed the theoretical merits of such an approach (e.g., Fattahi, 1995; Pattie & Cox, 1996; Drabenstott, 1996; Younger, 1997). In general, these studies have agreed that new metadata standards will need to move beyond the card-based model of MARC and traditional cataloguing systems. With these ideas in mind, in 1995 the OCLC (Online Computer Library Center) convened a meeting at Dublin, Ohio to propose a new "Dublin Core" standard for online cataloguing. (Weibel, Godby, Miller, & Daniel, 1995) During a sequence of some half-dozen conferences, the Dublin Core has emerged as a core set of some 15 descriptive elements (OCLC, 1997). Though the Dublin Core has not been adopted officially as a standard by international library organizations, it remains the best-documented proposal for such a standard. In the reorganization of online cataloguing for the IRN, I decided to base the format closely on the Dublin Core. Though this was not necessary from a technical perspective, I wanted to ensure long-term viability of cataloguing records by sticking with recognized naming conventions. This decision has been reinforced by the recent arrival of commercial search tools that automatically recognize and index Dublin Core tags. For example, Ultraseek, an intranet search engine by Infoseek, has adopted automatic indexing of Dublin Core metatags (Infoseek, 1999). Ultraseek can recognize such tags as “dc.title” and “dc.abstract” as official descriptions of documents.

In conjunction with the adoption of Dublin Core metadata, I decided to adopt the new Resource Description Framework (RDF) proposed standard of the World Wide Web Consortium (W3C). RDF is an attempt to define a universal standard language for describing metadata (Heery, 1998). Figure 1 illustrates this somewhat confusing hierarchy of standards. In practice, this hierarchy is simple to implement. RDF utilizes the general language syntax of XML (Extensible Markup Language), which is itself a radically simplified version of SGML, the grandfather of all markup languages. RDF offers a great improvement over HTML "Meta" tags for describing documents, by adding a structured syntax. When implemented in RDF, Dublin Core gains greatly in elegance (Powell, 1998). Figure 2 illustrates the simplicity of this format.

```
<rdf:RDF
   <rdf:Description about="mrr_981020.html">
     <dc:Title>US ADSL Market Forecast, 1998-2001</dc:Title>
     <dc:Description>The ADSL market stands poised for massive near-term growth.  This report
focuses on the likely areas of growth within the US market until 2001.</dc:Description>
     <dc:Creator>Seymour Blab</dc:Creator>
     <dc:Publisher>Market Researchers Anonymous</dc:Publisher>
     <dc:Date>1998-11-07</dc:Date>
   </rdf:Description>
  </rdf:RDF>
```

Figure 2: RDF Example Fragment

In practice, the Dublin Core implemented in RDF behaves like any other set of SGML-compliant tags. Because the IRN's commercial search engine supports the indexing of SGML tag sets, RDF was a practical option for the IRN organization. The details of this implementation will be discussed in the "methodology" section.

My survey of the available literature on metadata revealed a wealth of theory and a dearth of implementation studies (IFLA, 1999)[1]. Nearly all of these projects have been conducted in academic and government research settings, without much accompanying documentation. In addition, even the best documented examples have not utilized RDF syntax, and have relied on the old HTML "Meta" tag format (Powell, 1997; Hakala, 1998). Neither have these studies attempted to measure the effectiveness of these metatags for improving search performance. Hence my study of a metadata search engine at the IRN is unique in several aspects. First, I am documenting a large-scale implementation of Dublin Core metadata using RDF syntax. Second, this collection is located in a major corporate digital library. Third, an attempt has been made to measure the effectiveness of this implementation, based on feedback from trained corporate

[1] IFLA (International Federation of Library Associations and Institutions) maintains the most comprehensive and up-to-date bibliography of metadata resources. See http://ifla.inist.fr/II/metadata.htm.

research specialists. By implementing leading-edge standards in a production environment, I have faced some novel challenges. This study will document the effort to overcome these challenges with a mixture of technical and organizational methods. Together, these methods offer an intriguing example of information architecture in a corporate setting.

III. METHODOLOGY

Implementing the Metadata

The first step in this project was the creation of a collection of searchable, indexed metadata. With more than 30,000 online documents in a variety of formats from more than a dozen vendors, this presented quite a challenge. Two major issues presented themselves: where should we physically locate the metadata, and how would we combine metadata from multiple sources?

Metadata Location

In most available studies of metadata, it is customary to include the metadata within the source document, either as XML (RDF) tags or as HTML "Meta" tags. This solution allows for simple indexing of documents and unites content with metadata. However, it has two disadvantages. First, such metadata cannot be added to Adobe Acrobat (PDF) documents or other proprietary formats, including Microsoft Office. Second, in-document metadata can be difficult to update. By contrast, metadata in separate files can be manipulated independently from the source documents. One innovative approach involves keeping the metadata in separate files, and using web "server side includes" (SSI) to add the information as it is delivered by the web server (Powell, 1997). For the IRN, I chose a hybrid solution to this problem. Because the document collection includes hundreds of PDF and MS Office documents, it was necessary to create matching metadata files for those documents. In these cases, a matching filename was created with the extension ".rdf." Since the IRN's search engine bypasses the web server to access documents directly, the SSI method could not be used. Nor could the search engine recognize "paired" description files -- the engine only recognizes metadata for the document currently being indexed. Recognizing these limitations, I adopted an "in-document" metadata strategy for HTML documents, while using "paired descriptions" for the non-HTML documents. This solution is not optimal, because the search engine cannot index both the full-text and metadata for non-HTML documents. Such documents must be treated as a special category, due to the limitations of current search engine technology. Thus, in the IRN system, metadata is indexed and searched separately from source documents, with formats other than HTML.

Reconciling Multiple Metadata Sources

The IRN's cataloguing team had manually catalogued nearly 1,000 of some 30,000 available online documents using a traditional MARC-based OPAC. These OPAC records included title, date, abstract, keywords, and document URL. These URLs often pointed to a single table of contents for a multi-part report or series of reports. When these "table of contents" references were extrapolated, the real total of catalogued documents approached 5,000. Because the OPAC uses a proprietary data format, direct data conversion was not possible. Instead nightly text reports of new online materials are produced and parsed with Perl scripts to extract the latest catalog records. The Perl scripts convert this data into RDF records, which are inserted into the appropriate HTML or paired ".rdf" files.

For the remaining 25,000+ documents not covered in the OPAC records, we had to depend on vendor-supplied information. Some of the largest vendors already supplied HTML "meta" tags with their documents. In such cases, our Perl scripts simply translate the HTML metatags into their equivalent RDF tags. When both OPAC and vendor records existed, I combined the information into a single aggregate field. Our cataloguers felt that we should err on the side of completeness in such cases. When vendor-supplied metatags were not present, more creative solutions were necessary. In most cases, our Perl scripts parsed the vendor-supplied directory and file names for date and subject patterns. Such methods are a last resort, because market research vendors often change the directory structure and filenames of their deliverables, forcing us to rewrite the Perl scripts to deal with the changes. Figure 3 illustrates the end result of these data conversions. At a minimum, all documents include five fields of metadata: title, publication date, publisher, relation, and type. Subject and description are added when available.

<html>	
<head>	
<title>US ADSL Market Forecast</title>	
<rdf:RDF	Begin the RDF Tagset
xmlns:rdf="http://www.w3.org/TR/WD-rdf-syntax#"	Points to the official RDF syntax definition (an XML feature).
xmlns:dc="http://purl.org/metadata/dublin_core#">	Points to a similar definition for Dublin Core elements.
<rdf:Description about="mrr_981020.html">	Beginning of Description block, starting with optional URL
<dc:Title>US ADSL Market Forecast, 1998-2001</dc:Title>	Title of the source document.
<dc:Description>The ADSL market stands poised for massive near-term growth. This report focuses on the likely areas of growth within the US market until 2001.</dc:Description>	Abstract description of the source document.
<dc:Creator>Seymour Blab</dc:Creator>	Author of the source document.
<dc:Publisher>Market Researchers Anonymous</dc:Publisher>	Publisher of the source document.
<dc:Date>1998-11-07</dc:Date>	Date of publication. (yyyy-mm-dd)
<dc:Subject>ADSL, xDSL, Market Forecast, US Telecom </dc:Subject>	Subject description block.
<dc:Type>Market Research Report</dc:Type>	Conceptual document type
<dc:Format>text/html</dc:Format>	Document storage format -- main options are HTML and PDF
<dc:Identifier>MR-563094</dc:Identifier>	Unique identifier (optional)
<dc:Relation>	Sample Relationship tag. This document is a part of
<dc:RelationType="IsPartOf"/>	www.mrr.com/index.html.
<rdf:value resource=http://www.mrr.com/index.html />	
</dc:Relation>	
</rdf:Description>	End, Description block
</rdf:RDF>	End of RDF block
</head>	
<body>	
<P>This is an example.</P>	
</body>	
</html>	

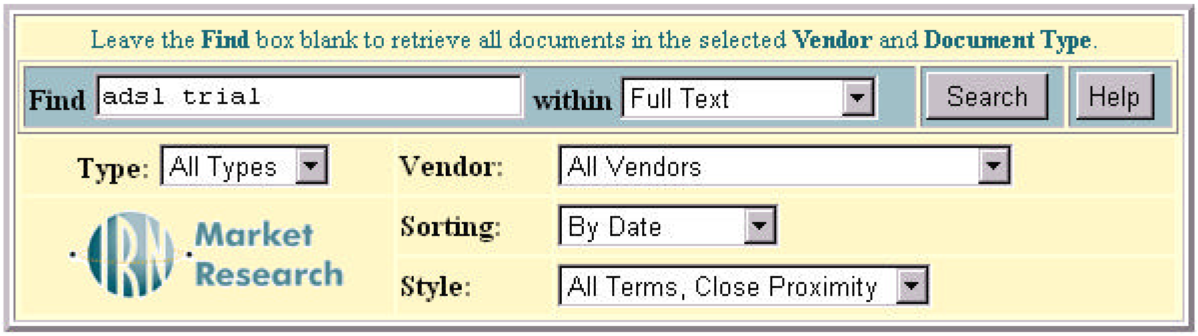
Figure 3: Sample RDF record inside an HTML file

Implementing the Search Interface

The IRN web site utilizes the commercial search engine Excalibur Retrievalware, version 6.6.2. Though Excalibur advertises Retrievalware's "concept-based" search abilities, specialists complained frequently about the results of such full-text searches. I decided to make use of Excalibur's lesser-known SGML and fielded searching capabilities. This proved to be easier than expected, after the difficult task of adding RDF metadata to all 30,000 documents. With a common metadata format available for all documents, Retrievalware only needed a single indexing tag definition file. Once the search libraries were re-indexed using the fielded metadata, I designed a search interface (see Figure 4). This interface was based on feedback from our users, who demanded four features:

1. Ability to search document titles.
2. Ability to limit searches by publisher.
3. Ability to sort by date.
4. Ability to search with Boolean expressions.

These features were all supported, along with additional catalog fields and search types.



Leave the **Find** box blank to retrieve all documents in the selected **Vendor** and **Document Type**.

Find: adsl trial within: Full Text

Type: All Types

Vendor: All Vendors

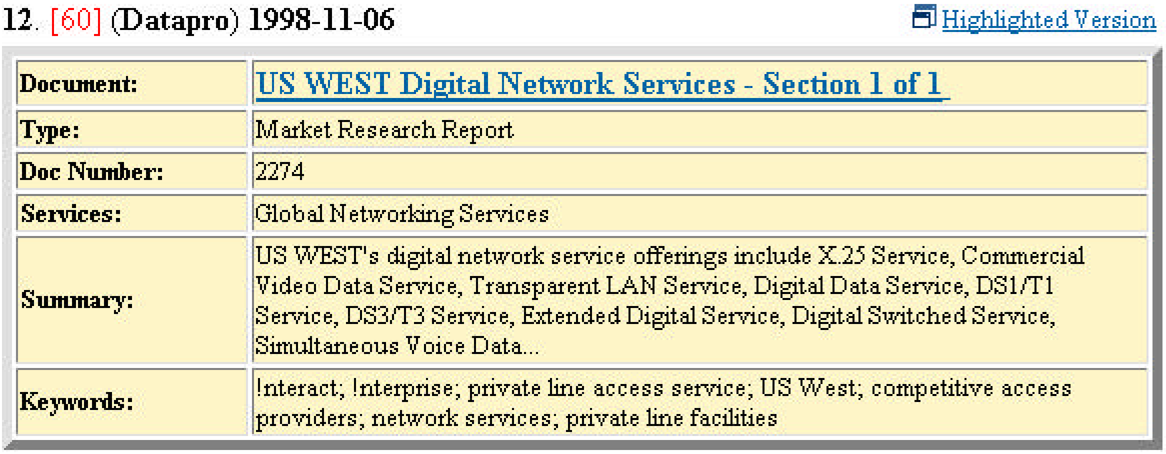
Sorting: By Date

Style: All Terms, Close Proximity

Figure 4: IRN Market Research Search Interface

The new search interface was officially released on December 1, 1998. Through the use of drop-down options, it was possible to include many options in a compact interface. Query terms can be sought in five different fields: "full text," "all catalog fields," "document title," "document author," and "document number." Searches can be limited by document type, using either "market reports" or "market briefs." Users may select one of seven major research vendors: IDC, Datapro, Datamonitor, Yankee, Meta Group, Forrester, and Financial Times. Results may be sorted by date, or by relevance ranking. Finally, an innovative option allows for five search styles: "All Terms, Close Proximity," "All Terms, Anywhere," "Any Terms," "Any Terms, and Synonyms," and "Boolean." We had received many complaints from specialists about Excalibur's ill-defined "concept" search mode, along with its inability to rank Boolean query results. With these five styles, I attempted to make search methods self-explanatory, rather than simply offering Excalibur's two default "concept" and "Boolean" options.

Behind the scenes, Javascript code translates these abstract styles into concrete Excalibur search expressions. All queries are run as "concept" mode to produce relevance rankings, but they are filtered using Boolean expressions. For example, a close proximity search for "ADSL trial" is translated as follows: concept mode "ADSL trial", with a Boolean filter for "ADSL trial within 20." Excalibur performs a simple "Or" search with the concept mode, and then removes all results that do not match the Boolean filter. By creating these five self-explanatory search styles, I insulated users from an unnecessary degree of search syntax complexity.



12. [60] (Datapro) 1998-11-06 Highlighted Version

Document:	US WEST Digital Network Services - Section 1 of 1
Type:	Market Research Report
Doc Number:	2274
Services:	Global Networking Services
Summary:	US WEST's digital network service offerings include X.25 Service, Commercial Video Data Service, Transparent LAN Service, Digital Data Service, DS1/T1 Service, DS3/T3 Service, Extended Digital Service, Digital Switched Service, Simultaneous Voice Data...
Keywords:	!nteract; !nterprise; private line access service; US West; competitive access providers; network services; private line facilities

Figure 5: IRN Result List Sample

Beyond customizing the initial search interface, I also created a "library card" format for search results (see Figure 5). For each document in the result list, all available metadata fields are displayed. Excalibur allows for a "highlighted version" of each document, where the query terms are outlined in bright colors. This library card is reproduced at the top of each highlighted document, so that searchers can print the catalog information along with the full document. This display format resulted from complaints about the out-of-box result lists, which only displayed document title and relevance ranking. Research specialists felt that they could not determine document relevance from so little information. With the library card result list format, I eliminated this complaint by providing a metadata-rich result list.

Evaluation of IR Tool Performance

Survey Design

In order to determine the effectiveness of the new IRN search engine, I decided to implement a survey of library staff. Because these employees are located in seven locations worldwide, from Hong Kong to Britain, web-based evaluation methods were deemed necessary. Several months were allowed to elapse, so that the library staff could become accustomed to their new search tool. Because I intended to ask about average usage patterns for various search tools including the new search engine, it was important to allow such a period for acclimatization. For business reasons, it was necessary for the survey to ask about all available search tools, rather than dealing exclusively with the IRN's internal search engine. Department-wide surveys are infrequent, and therefore tend to be rather broad in scope. Due to these requirements, I adopted a strategy of comparative search engine evaluation. Questions would aim at the general search experience of participants with various tools, but would focus on using these other tools as a baseline for comparison with the IRN's internal search engine.

In designing the survey questionnaire, I followed the methodology suggested in Babbie's The Practice of Social Research (Babbie, 1995). For constructive criticism, I consulted Dr. Diane Sonnenwald and Dr. Evelyn Daniel, professors at the School of Information and Library Science at the University of North Carolina. In addition, I reviewed the questions with Beverly Wiggins, Associate Director of the Institute for Research in Social Science at UNC. During the final construction of the survey, I pre-tested the questionnaire with four IRN staff members at the corporate library. I incorporated many recommendations and comments from these reviewers in the final version of the survey. These comments were especially helpful for ensuring consistency and clarity of terminology in the questionnaire. The final survey was reviewed and approved by the UNC-CH Institutional Review Board.

The survey included a total of 56 questions, divided into seven HTML pages for an average of eight questions per page. The entire survey is shown in Appendix A. Nine of the survey questions were open-ended, and the remainder were closed-ended. Much attention was paid to the order of the survey questions. The first two pages, with 14 questions, aimed at establishing the search experience and general job function of the participant. Participants were asked about their average weekly online search workload, along with their usage of various available search tools. They were also asked to estimate their level of expertise with these same tools. By separating usage from estimated expertise, I aimed to determine whether certain tools were especially difficult to learn. The questions about job function offer another comparative baseline between IRN staff, so that I can identify distinct user groups within the 90-person staff who have unique search requirements.

The next three pages of the survey focused on in-depth evaluations of three major types of search tools: Internet Search Tools, Commercial Search Tools (i.e. Dialog, Lexis/Nexis), and the IRN Search Engine. Originally I planned to ask about the precision and accuracy of these tools, with ratings using a Likert scale of percentage ranges. However, my pre-test indicated that participants would not know how to answer these questions. Without a detailed side-by-side comparison of search result lists from similar searches, such estimates of performance would be hopeless. In addition, those figures would not have been especially valuable, because the search domains were so different. An Internet search engine might well search 20 Million pages of unknown origin, while the IRN Search Engine would search 30,000 documents in a limited subject domain. Instead of comparing these search tools by objective performance criteria, I decided to compare them by features and popularity. Which Internet and commercial tools were used most frequently by IRN staff, and why? What features from these popular tools might be used for the IRN search engine? For the closed-ended feature ratings section, I chose features that were readily comparable between all three types of tools. These features include result rankings, result sorting, result display, search syntax, and category searching.

During the section comparing the three types of search tools, participants proceed from the most general type of tool (Internet) to the most specific (IRN Internal). With all of these tools, an option is available to open the tool in a separate window as a "memory refresher" during the survey. I chose this ordering for two reasons. First, survey participants would be familiar with the ranking system by the time they reached the IRN search engine. Second, participants would understand that they were making comparative rankings. All the open-ended questions about features explicitly refer to the IRN search engine: "What feature most distinguishes this product from other (Internet) search tools? Why might this feature be incorporated in the IRN search tool?" For the IRN search tool itself, I asked a minor variant: "What feature(s) most distinguish the IRN Web Site Search Engine from similar site search tools?" In addition, "How would you improve the IRN Web Site Search Engine? What additional capabilities would you add?" These questions intentionally overlap, following the survey principle that one should ask the most important questions in several different ways. In this manner, it is possible to spark a valuable response at multiple points in the survey.

Page six of the survey focused entirely on fielded category searching, the real meat of the metadata innovations. Participants are asked how frequently they search with the category fields; next they were asked to rate the value of specific fields for their work. All of the available fields from the IRN Market Research interface were included in this list: catalogue fields, document title, document author, document number, document type, publisher, and service name. "Value" is an important variable here, because it asks users to rate the perceived value of a category in relation to their everyday work, independently from the frequency with which they actually use these fields. An open-ended question asks, “What other searchable fields would be valuable for your work?” By this point in the survey, I planned for participants to have their minds fully on the IRN search engine. This page is the real focal point, when it becomes clear that participants are evaluating a specific tool in great detail. I was guided by the principles of optimal survey design, proceeding from the most general to the most specific questions.

The final survey page aimed to evaluate the desirability of possible future features, many of them at least potentially related to metadata. These options include topical search trees, personal search agents, document excerpts in results, search refinement, query by example, and multiple field sorting. Finally, two questions deal with types of information sought with search tools. First, “What types of information do you search most frequently for your job?” Second, “What types of information need improved coverage by the IRN search engine?” These questions are closed-ended, but they offer an open-ended “other” option and allow for multiple selections. Hence the final page focused on possibilities for future improvements to the search engine, both with new features and new areas of coverage.

Survey Distribution

The survey was designed and distributed using a corporate-wide tool known as "Surveycom." Surveycom utilizes a Java application for designing web-based surveys. Survey results are kept in an Oracle database, accessible using Cold Fusion. Surveycom offers a variety of options for internal corporate surveys, including a range of open-ended and closed-ended question types. Survey participants are anonymous, and no tracking information is provided for survey designers. Surveycom compiles survey results in its database, and allows these results to be downloaded in tab-delimited spreadsheet form. This format is imported easily into major spreadsheets and statistical programs such as MS-Excel and SPSS. In addition, Surveycom produces graphs of total survey results for quick browsing in HTML form. The survey and its accompanying graphs are included in the appendix.

Once I created the survey using Surveycom, distribution was a simple matter of pointing participants to a URL with a web browser. I distributed a cover letter by e-mail, explaining the rights of participants and the reasons for the survey. I allowed a two-week period for survey responses, from March 17 to April 1. A prize was offered for "Early Bird" respondents during the first week, and a second prize was offered for all participants. These prizes (a Dilbert book and a $25 gift certificate) were provided by the department, and were consistent with standard departmental survey procedures. In addition, the Institutional Review Board at UNC-Chapel Hill approved this methodology. On March 17, I e-mailed the cover letter to 81 members of the IRN, officially launching the survey.

Log Analysis

Beyond the surveys, I utilized a second method of analysis for measuring the overall effectiveness of changes to the IRN search engine: log file analysis. Since November 1998, the IRN search engine has recorded log files of search behavior. These log files include query terms, query fields, and number of search results. No identifying information is included in these logs. With these log files, it is possible to measure the frequency of usage for the field-based query options since they were introduced in December 1998. The number of search results from these queries can also be analyzed for significant patterns. Unfortunately these log files are limited, because it is impossible to identify continuity of "sessions," or to determine how many query results were actually examined by a user. Hence these logs provide some valuable information about aggregate search behavior, but do not allow deep analysis of individual queries.

IV. DATA ANALYSIS

Survey Returns

The survey generated a good response rate from IRN searchers. Participants chose their role in the IRN from five categories: “information specialist,” “administrator,” “manager,” “technical specialist,” and “associate specialist.” IRN technical specialists support the department's computer systems. Information specialists are the largest single group within the IRN, with 28 staff members in this category. These specialists were the prime audience for the survey, because they are the professional researchers who answer most reference questions. 18 information specialists responded, for a 64.3% response rate. The IRN’s associate specialists were the next largest target group for the survey, since they represent the “first line” of reference at the IRN Call Center. The Call Center is a 24 hour/weekday phone service, available worldwide for corporate employees. When associate specialists at the Call Center cannot answer a question immediately, they forward the query to an information specialist. Seven of ten associate specialists responded to the survey, for a 70% response rate. Overall, 66.8% of the IRN’s 38 professional researchers responded to the survey.

By contrast, the response from non-searchers was less impressive. There were three respondents from each of the three remaining categories: “administrator,” “manager,” and “technical specialist.” These responses totaled 24.3% of 37 non-searchers. I expected a lower response rate from these groups, because they do not perform online searching as a major part of their job descriptions. For statistical purposes, I have combined these three groups into a single “administrative” category. Henceforth I will discuss three statistically distinct user groups: information specialists, associate specialist (or associates), and administrative. All the surveys were completed in their entirety, for closed-ended questions. The open-ended questions generated approximately an average 48% response rate.

Overall I was impressed by the volume and quality of responses. The majority of trained searchers responded enthusiastically, with the first 28 responses arriving during the first two days of the survey. The remaining six surveys arrived on the eighth day of the survey, immediately following an e-mail reminder message. It appears that Surveycom is a highly effective method of survey distribution, judging from the extreme rapidity of responses. This web-based method appears to avoid the traditional disadvantages of paper-based survey rates, known for their long response time and low return rate. Though I would have liked more responses from the "administrative" staff at the IRN, this pattern was predictable. When searching is not a major part of one's job description, a survey of search tools might seem irrelevant. Fortunately the nine responses from this group are statistically significant, and offer an excellent "control group" for comparison with the professional searchers. All these survey results were recorded in tab-delimited spreadsheet format by Surveycom, and imported directly into SPSS version 8.01, a statistical analysis software package for the social sciences.

Search Frequency of Respondents

In order to determine the search patterns of IRN staff, I asked "How many hours of online/web-based searching do you perform in a typical work week?" The question utilized an ordinal scale, with ranges in increments of five hours. Results confirmed my decision to divide the users into three major groups. Information specialists averaged a mean of 19.7 hours searching per week, with a minimum of 5-10 hours of searching per week. Four of these specialists spend 26-30 hours searching per week, easily the majority of their working hours. Associates returned a mean of 14.4 search hours, ranging from 3 to 28 hours with a 9.0 hour standard deviation. Administrative staff averaged only 5.4 hours of weekly searching, without a single employee spending more than 5-10 hours searching. Within this group, managers spent the least time searching , with a mean of only 3.7 hours. These numbers confirmed my hypothesis about the distribution of searching activities within the IRN. Clearly the information specialists use the search tools most frequently, while the associate specialists run a close second. Administrative staff are quite distinct from the others, performing comparatively few searches.

	Dialog	Internet Tools	Corporate AltaVista	Vendor Tools	IRN Infocat OPAC	IRN Search Engine
Info. Specialists	5.3	9.7	5.1	10.1	9.1	9.2
Assoc. Specialists	1.7	4.4	3.9	4.4	12.3	7.3
Administrative	0	10.2	2.8	3.2	5.0	4.9

Figure 6: Mean Search Frequencies by Role and Tool

In questions about search frequency with individual tools, I hoped to differentiate tool usage between different user groups. Survey results concerning search frequency with available tools are illustrated in Figure 6. Dialog showed the greatest variation between groups. The average information specialist performed 5.3 searches per week with Dialog, while associates averaged 1.7 searches and administrators averaged zero. Administrative staff averaged less than five searches per week with all six tools, except for the dramatic exception of Internet searching. Administrators averaged 10.2 Internet searches per week, compared to 9.7 for specialists and 4.4 for associates. The corporation’s internal search tool was the least-used tool, with a maximum average frequency of 5.1 searches/week for information specialists. This is an important figure, because the corporation's internal AltaVista is the main avenue for finding internal corporate documents. As we will see in later questions, IRN staff consistently demanded better coverage of this type of information.

With the three tools focusing on market research (Vendor tools, IRN Infocat OPAC, and IRN Search Engine), survey results were revealing. With vendor-specific search tools, information specialists averaged some 10.1 searches per week, the second-highest frequency for any tool. Meanwhile, associates averaged less than five searches with vendor-specific tools, and administrators less than 4. By contrast, associates were the heaviest users of the Infocat online OPAC, with some 12.3 searches per week. Information specialists averaged 9.1 searches with Infocat. On the other hand, those specialists averaged 9.2 searches with the IRN Search Engine, while associates averaged 7.3 searches. For all of these tools, administrators averaged slightly less than five searches per week.

In summary, search frequency data reveals distinct tool usage patterns between the three major IRN user groups. Information specialists were the most frequent users, across the board. They were the only significant users of Dialog and Vendor-Specific tools, and they were the heaviest users of the IRN search engine and corporate AltaVista. Such heavy and varied tool usage is consistent with the role of information specialists, who must track down research requests wherever the information can be found. Associate specialists were surprising in their heavy reliance on two tools: the IRN Infocat OPAC and the IRN search engine. For the other tools, they barely averaged more than administrative staff. This pattern probably results from the Associates' work with book circulation, a role which is rapidly vanishing as the IRN eliminates its book collection and goes all-digital in April 1999. Associates' reliance on these two tools indicates that they are mainly concerned with materials held by the IRN, both online and hardcopy. When questions cannot be answered from these internal sources, Associates will often refer the query to an information specialist. Finally we reach the administrative staff, who ranked at the bottom with all tools except Internet searching. Clearly these users have different information needs, not needing to focus on answering client queries. Hence they rely on external Internet sources for personal discovery of information outside the corporation. Administrators were also relatively significant users of the IRN's internal search tools, indicating a secondary interest in market research materials.

Search Expertise of Respondents

	Dialog	Internet Tools	Internal AltaVista	Vendor Tools	IRN Infocat OPAC	IRN Search Engine
Info. Specialists	4 (1-5)	4.2 (3-5)	3.9 (1-5)	4.1 (2-5)	4.4 (3-5)	3.6 (2-5)
Assoc. Specialists	2.1 (1-3)	3.7 (3-5)	3.4 (3-4)	2.7 (2-4)	4.3 (3-5)	2.7 (2-4)
Administrative	1.8 (1-5)	3.7 (2-5)	3.1 (2-4)	3.1 (2-4)	2.7 (1-5)	3.0 (1-5)

Figure 7: Mean Search Expertise by Role and Tool

(5-point scale, 1=None, 2=Novice, 3=Comfortable, 4=Experienced, 5=Expert)
Range of values is indicated in parentheses.

Compared with search frequencies for the various tools, claims of search expertise revealed some disparities. Once again, information specialists were strong across the boards, averaging "Experienced" with all tools. They registered the least expertise with the IRN search engine, with a figure of 3.6. Administrative users averaged the lowest, feeling "comfortable" with all the tools except Dialog, which they never use. Associates were the biggest surprise, with averages similar to administrators for any tool other than the IRN OPAC. Because they are the heaviest users of the Infocat OPAC, their 4.3 "Experienced" average was no surprise. However, they are the second heaviest users of the IRN search engine, but they registered only a 2.7 expertise with this tool, the lowest of all three groups. Expertise with the IRN search engine appears to be a serious issue, because its most frequent users do not express great confidence in their understanding of the tool. This trend is evident with both information specialists and associates. Another surprise were the administrators, who ranked slightly higher than associates for their expertise with vendor tools and the IRN search engine.

Implications

Survey results on search expertise clarified the important distinctions between IRN user groups. Information specialists consider themselves masters of all search tools, and seem slightly puzzled with some of the idiosyncrasies of the new IRN search engine. Hence their slightly lower “expertise” rating with the IRN search engine. We will see their detailed thoughts in a later section, when they offer their open-ended thoughts on the tools. The expertise patterns of associate specialists are slightly more difficult to interpret. They are very strong in some areas, and express little confidence in others. This probably results from an emphasis on training associate specialists, as the “first line” of response. The IRN defines the roles of associate specialists rather strictly, and insists on training them with any new tools before “officially” allowing them to use these tools. No such formal training program has been offered for the IRN search engine, which probably explains the low expertise ratings for associates with this tool. Administrative personnel, often with a more technical background, seem more comfortable learning new tools on the fly, but do not use these tools frequently enough to consider themselves “experts.” Managers and technical employees appear to express more confidence in their expertise, despite using search tools less frequently than other groups. Clearly these are distinct user groups with different needs and attitudes toward searching. As we will discuss later in the paper, future plans must tailor themselves for these different groups. The IRN search interface may need to change to meet their preferences, while more training may be helpful for people using the internal search tools.

Favorite Internet Search Tools

	AltaVista	Yahoo	Hotbot	Infoseek	Metacrawler	Northern Light	Excite	Total
Info. Specialists	9	2	3	2	2			17
Assoc. Specialists	4	1				1	1	7
Administrative	4	3				1	1	9
Total	17 (50%)	6 (17.7%)	3 (8.8%)	2 (5.9%)	2 (5.9%)	2 (5.9%)	2 (5.9%)	34 (100%)

Figure 8: Most Frequently Used Internet Search Tools, by Role and Tool
(Total percentage indicated in parentheses)

Among Internet search tools, AltaVista emerged as the clear favorite, with 50% of the total vote. Yahoo ran a distant second, with 6 users for 17.7% of the results. Both of these tools were popular among all user groups. No other tool garnered more than 3 votes. Three tools were notable for their relative popularity among information specialists: Hotbot, Infoseek, and Metacrawler.

	AltaVista	Yahoo	Hotbot	Infoseek	Metacrawler	Northern Light	Excite	Avg.
Meaningful Rankings	3.8	3.2	4.3	4.0	2	4.5	3.5	3.6
Result Sorting	3.2	*	3.0	*	2	4.5	2.5	3.1
Display of Results	3.3	3.0	3.7	4.0	3.5	4.0	2.5	3.3
Search Syntax	3.6	*	3.7	4.0	3.5	5	3.0	3.5
Category Searching	3.0	*	4.3	4.0	1	4.5	2.5	3.2

Figure 9: Mean Ratings of Internet Search Tools, by Feature and Tool

(5-point scale, 1=Unacceptable, 2=Mediocre, 3=Acceptable, 4=Good, 5=Outstanding)
* indicates most or all users answered "n/a" for this category.

Comparing the rankings of individual features for Internet search tools, a surprising pattern emerges. Users did not rank the most popular tools especially highly. AltaVista rated best for its ranking and search syntax (3.8 and 3.6), but its other features scored little better than average. Yahoo scored beneath average for its ranking and result display, but was not ranked in several categories because of its unique nature as a topic tree.   Hotbot and Northern Light emerged as the clear favorites for overall features, but these ratings are less reliable due to the small sample size for these tools.

Why are the most popular tools rated so badly for individual features?   Some answers are found with responses to the open-ended question, "What feature most distinguishes this product from other Internet Search tools? Why might this feature be incorporated in an IRN search tool?" AltaVista was cited by several users for the "reliability and comprehensiveness" of its results, along with its strong support for Boolean searching. AltaVista's average scores are explained largely by the survey's focus on search features, rather than completeness of results. Likewise, Yahoo earned praise for its easily referenced topical "directory structure," a feature that was not included in the numerical rankings. Hotbot and Northern Light were notable for the devotion of their users, who poured compliments on these relatively new search engines. Hotbot received acclaim for its strong field-searching capabilities, along with its ability to refine searches. In addition, two users singled out Hotbot's precision, with its ability to show the "top 10 most visited sites for a given search string." On the other hand, Northern Light earned its popularity with a unique "folder tree view of results" which greatly impressed its users. These tools appear to be gaining in popularity, but are handicapped by their relative youth and a perception that their more established counterparts can deliver greater completeness of results. Hotbot was especially notable for its niche popularity among information specialists, the most frequent IRN searchers.

Implications

A single overwhelming lesson emerges from the ratings of Internet search tools: in the end, the popularity of a tool results from its ability to deliver high completeness and precision of results. These factors create a feeling of trust for the tool, which is the basis for further searching. It was striking to see the relative unpopularity of the Internet search tools with the most highly rated features, such as Hotbot, Infoseek, and Northern Light. These discrepancies can only be explained by lack of trust, perhaps compounded by the relative youth of these products. The most popular search engine, AltaVista, earns its ranking by advertising the largest index of Internet materials, accessible with a straightforward Boolean search strategy. AltaVista does not claim to have the flashiest features; instead it promises reliability and completeness. Yahoo offers similar virtues – nobody claims that its search mechanism is especially feature-rich. Rather, Yahoo earns its popularity with its library-style hierarchy of categories, maintained manually by human cataloguers. Users of Yahoo are not looking for exhaustive completeness of results, rather they are seeking human judgements on the *best* sites in a category. Hence the synergy between Yahoo and AltaVista is a brilliant marketing move – "if you can't find it on Yahoo, we'll send you to AltaVista."

The IRN can learn from these examples. With metadata I have created many search options that were previously impossible. However, survey respondents indicate that we must work harder to build that bond of trust in the reliability of the IRN search engine. Searchers do not care if they can search by field, if they do not believe that they will receive a complete list of results.

Likewise, there is a widespread desire for a Yahoo-style category search option, integrated with the current search engine. With a foundation of metadata that describes the characteristics of each document, it should be much easier to build such a category structure in the future. With self-describing documents, the hierarchy can effectively "build itself," without needing extensive manual intervention. Automatic algorithms can parse the documents for metadata, and determine where they fit within a category structure. This could be a crucial benefit of metadata for online site architecture.

Favorite Commercial Search Tools

	Dialog	Dow Jones	IDC-Net	Computer Select	Reuters	Collectanea	IHS Standards	N/A	Total
Info. Specialists	11	3	3				1		17
Assoc. Specialists	1	5			1				7
Administrative		4	1	1		1		2	9
Total	12	12	4	1 (2.98%)	1 (2.9%)	1 (2.9%)	1 (2.9%)	2	34
	(35%)	(35%)	(12%)					(5.9%)	(100%)

Figure 10: Most Frequently Used Commercial Search Tools, by Role and Tool
(Total percentage indicated in parentheses)

Among commercial search tools, Dialog and Dow Jones tied for greatest popularity among IRN searchers, each garnering 12 votes for a total of 70% of all respondents. IDC-Net ran a distant third, with 4 results. No other commercial product earned more than a single vote, so that ratings for these products were statistically insignificant. 65% of information specialists chose Dialog as their favorite tool, with only a single associate specialist making this selection. By contrast, Dow Jones was extremely popular among associates and administrators, respectively chosen by 71% and 44% of those two groups.

	Dialog	Dow Jones	IDC-Net	Average for All Tools
Meaningful Rankings	4.3	3.8	3.8	3.9
Result Sorting	4.4	3.9	4	4.2
Display of Results	3.8	3.5	4	3.7
Search Syntax	4.9	3.7	2.7	4.0
Category Searching	4.9	3.9	4	4.1

Figure 11: Mean Ratings of Commercial Search Tools, by Feature and Tool
(5-point scale, 1=Unacceptable, 2=Mediocre, 3=Acceptable, 4=Good, 5=Outstanding)

Unlike the Internet search tools, a clear winner emerged in the ratings of commercial search tool features. Dialog earned nearly perfect marks from its users, with a 4.9 in both search syntax and category searching. Likewise, Dialog's ratings for ranking and result sorting were excellent. These ratings were the best of any search tool, and much higher than the average values for either commercial or Internet search tools. Dow Jones and IDC-Net both earned good rankings for search features, with the exception of IDC's search syntax. Overall, these ratings indicate that commercial search tools tend to contain more advanced search features than their Internet counterparts. This discrepancy can be explained by the fact that these commercial search engines index proprietary databases, which allow for search features that are impossible with an enormous unstructured collection such as the Internet.

In response to the open-ended question about the most valuable features of commercial search tools, participants offered a detailed glimpse of their search needs. All respondents agreed that

Dialog offers unparalleled ability to "search across as many as several hundred databases at once with a single strategy." In addition, Dialog offers "flexibility in using search operators and proximity operators, ability to search by field, and ability to combine and reuse search sets." With this combination of complex search operators and enormous breadth of coverage, Dialog emerged as the clear favorite of information specialists, the most frequent IRN searchers. On the other hand, Dow Jones earned its popularity for its coverage of late-breaking news stories. Respondents also praised its simple search interface, with good ability to sort results and refine searches. Meanwhile IDC-Net garnered acclaim for its ability to "maneuver through related documents" by finding other documents with the same key terms, in the manner of subject cards in traditional library catalogs. Overall, these three tools filled different needs for various employees. Information specialists preferred Dialog's breadth of coverage and powerful search operators. IDC-Net was popular with others who needed more specific access to market research reports. Dow Jones was the clear favorite of associates and administrators, who tend to follow current news more than they perform exhaustive research.

Implications

The commercial search tools offer valuable lessons about metadata-based searching. All three of the products we discussed (Dialog, Dow Jones, and IDC-Net) are search tools for proprietary databases of documents, united by a common metadata format. Dialog and Dow Jones store their metadata separately from the documents, in a database. IDC-Net places metadata within their documents, but also appears to duplicate this information in a database. By using this database approach with a rigorous cataloging system, these products offer extremely advanced search functionality. Of course, all this functionality comes at a price. All of these services charge high fees which are beyond the reach of the general public. Corporate libraries, such as the IRN, buy large yearly contracts for these services because of the quality and reliability of their information coverage. It would be simply impossible to collect so much information as an in-house operation within the corporation. Hence a sort of “information outsourcing” is practiced by the IRN.

For the IRN’s internal search tool to compare favorably with these commercial tools, major resources must be devoted to the task. Currently the IRN search engine searches metadata wherever it can be found. Much of this metadata is supplied by more than a dozen different vendors, who use different formats and keywords for their document descriptions. Many vendors do not supply any metadata. In all these cases, the IRN’s cataloguing team should consider urging vendors to use standard metadata, or should increase its role with online materials. Traditionally, the cataloguers have dealt almost exclusively with hardcopy materials. Suddenly they have a new role, with the elimination of books from the IRN and an increase in digital documents. The IRN’s cataloguers will need to work very hard to establish a universal descriptive vocabulary for online documents, and to implement this vocabulary with metadata for online materials. They should aim for completeness and consistency of cataloguing records, so that the IRN’s search engine can perform on a par with comparable commercial products. The search engine is only as good as the documents it indexes.

Feature Ratings for the IRN Search Engine

The survey was designed to measure the effectiveness of the IRN search engine. By rating a variety of comparable Internet and commercial tools in previous questions, I hoped to establish a baseline for comparison of features (see Figure 12).

	Dialog	Dow Jones	AltaVista	New IRN Search Engine	Internet Average	Commercial Average
Meaningful Rankings	4.3	3.8	3.8	3.2	3.6	3.9
Result Sorting	4.4	3.9	3.2	3.4	3.1	4.2
Display of Results	3.8	3.5	3.3	3.3	3.3	3.7
Search Syntax	4.9	3.7	3.6	3.0	3.5	4.0
Category Searching	4.9	3.9	3.0	3.2	3.2	4.1

Figure 12: Comparisons of Most Popular Search Tools, by Feature and Tool

(5-point scale, 1=Unacceptable, 2=Mediocre, 3=Acceptable, 4=Good, 5=Outstanding)

Clearly the IRN search engine ranked lower than comparable commercial and Internet tools, in most categories. The IRN search engine's strongest feature was its result sorting, and its weakest feature was search syntax. Though these ratings were not spectacular, the IRN search engine scored better than "adequate" in all categories except search syntax. This was not surprising, because searchers have complained frequently about the IRN search engine's quirks in supporting certain Boolean search operators, such as "within" and "adj." Further, the original version of the engine did not allow Boolean search results to be ranked. Additional implications of these ratings will be discussed shortly.

Examining the breakdown of these ratings by IRN role, information specialists were the most critical users of the tool (see Figure 13). In every category, information specialists ranked Excalibur slightly lower than the average. On the other hand, administrators and associates ranked the tool significantly higher for result sorting, display of results, and search syntax. An analysis of search frequency shows a slightly different story (see Figure 14). For those who searched with the IRN search engine more than 15 times per week, the ratings of most features were significantly higher. Favorites of frequent searchers were result sorting (3.8 vs. 3.4 average), category searching (3.7 vs. 3.2 average), and meaningful rankings (3.6 vs. 3.2 average).

Strangely, the least frequent searchers expressed the greatest satisfaction with the display of results (3.6 vs. 3.3 average). This pattern indicates that the most frequent searchers would prefer to have a more advanced display of results, suitable for their specialized needs.

	Meaningful Rankings	Result Sorting	Display of Results	Search Syntax	Category Searching
Info. Specialists	2.9 (2-5)	3.2 (2-5)	3.1 (2-5)	2.7 (1-4)	3.0 (1-5)
Assoc. Specialists	3.3 (2-4)	3.7 (2-5)	3.2 (2-4)	3.2 (2-4)	3.7 (2-5)
Administrative	3.5 (2-4)	3.6 (2-5)	3.6 (2-5)	3.5 (3-4)	3.3 (1-5)
Average	3.2 (2-5)	3.4 (2-5)	3.3 (2-5)	3.0 (1-4)	3.2 (1-5)

Figure 13: Mean Ratings of IRN Search Engine Features, by Role and Tool

Range of values is indicated in parentheses.
(5-point scale, 1=Unacceptable, 2=Mediocre, 3=Acceptable, 4=Good, 5=Outstanding)

Weekly Search Frequency	Meaningful Rankings	Result Sorting	Display of Results	Search Syntax	Category Searching
<5 (12)	3.2	3.5	3.6	3.0	2.9
5-10 (11)	2.9	3.1	3.0	2.8	3.1
15+ (8)	3.6	3.8	3.1	3.2	3.7
Average	3.2	3.4	3.3	3.0	3.2

Figure 14: Mean Ratings of IRN Search Engine Features, by Feature and Frequency of Searching

(5-point scale, 1=Unacceptable, 2=Mediocre, 3=Acceptable, 4=Good, 5=Outstanding)
Number of participants in each category is indicated in parentheses.

Discrepancies between roles can be isolated by analysis of the open-ended questions about the IRN search engine's most valuable features, and about areas for improvement. The feelings of information specialists are best summarized by this quote, "I still need to increase my confidence in the search engine; the improvements of late are fantastic. It's just difficult to let go of the uncertainty… did it pull up everything?" Information specialists appear to have the greatest need for completeness of results, and they commonly voiced concern about this uncertainty. Much of this doubt may linger from the time prior to December 1998, when specialists expressed strong dissatisfaction with the older version of the IRN search engine. Many of them used the corporate AltaVista to search IRN materials, rather than using the IRN's own internal tool. Judging from search frequency data, this situation has been largely remedied. The IRN search engine is searched nearly twice as often as the corporate AltaVista (see Figure 6). However, the open-ended responses reveal that there is still room for improvement.

Respondents were quite willing to offer ideas for such improvements. Information specialists were the most vocal, with 61% offering detailed observations. By comparison, 44% of administrators and 14% of associates offered open-ended responses. The following list summarizes their demands, ranked by frequency of comments:

1. Clarity about which materials are searched, and not searched. Currently the IRN search engine searches market research, but not standards. In addition, it does not index some of the older market research materials. Clients want this coverage expanded and clarified.
2. Improved relevance ranking. This has always been a point of criticism with the IRN search engine, because it does not "weight" documents based on percentage of hits. A document with a single "hit" will rank the same as one with a dozen. This problem is unavoidable due to the engine's proprietary ranking mechanism, and is a strong reason for examining alternative search products.
3. "More ability to search by field and use search and proximity operators - more like Dialog or the IRN catalogue." Dialog is frequently held as a model for search syntax and field searching. While users are happy with the IRN search engine's new fielded search capabilities, they want more options. Some suggestions include: limiting by date range, and adding a keyword field. Keywords should conform to a consistent vocabulary.
4. "The 'help' manual should explain its search facilities more clearly with examples."
5. Search refinement should be more explicitly supported, allowing the reuse of result sets.

These ideas clearly demonstrate the reasons for the IRN search engine's relatively average feature ratings. Though IRN staff express satisfaction with the recent changes, they demand additional improvements to make this tool a more valuable resource. For the IRN search engine to achieve comparable ratings with the best commercial and Internet search tools, it must offer a comparable level of search features, completeness, and reliability. Judging from open-ended comments, this goal must still be reached.

Implications

The IRN search engine was ranked reasonably well by IRN staff for its features. However, they pointed out a great many areas for improvement. Most of these issues were not directly related to metadata, except for the request for more field searching options with a new interface allowing for searching multiple fields simultaneously. Instead criticism focused on completeness of results, and improved relevance ranking. Searchers want to know exactly which materials are being searched, and which ones are *not* being searched. Though they like the new date sorting options, they continue to be disappointed with the IRN search engine's relevance ranking algorithm. Documents are not "weighted" by hit density, as one would expect from common experience with search engines. Normally a document is ranked highly if it contains large numbers of "hits" on the query terms. The IRN search engine does not care about frequency of query term occurrence, a truly frustrating quirk for everyone involved. One searcher suggested that the IRN should throw away the current search engine and find a new search tool. This option is being considered by the IRN technical staff.

Meanwhile, metadata provides a solid foundation for the future, regardless of the tool chosen. So long as the documents are described in a standard metadata format, such as the IRN's RDF metadata scheme, a wide variety of search tools and databases can make use of this information. In short, it would appear that the addition of metadata has dramatically improved the IRN's internal search capabilities, but that users are still dissatisfied with the idiosyncrasies of the IRN search engine. I will elaborate on this distinction in the next section.

Ratings of Fielded Search Options

	Info. Specialist	Assoc. Specialists	Administrative
Search Frequency	3.6 (1-6)	3.4 (1-5)	2.3 (1-6)

Figure 15: Mean Frequency of IRN Field Search, by Role and Frequency

(6-point scale, 1=Never, 2=10%, 3=25%, 4=50%, 5=75%, 6=Always) Range of values is indicated in parentheses.

With usage of the IRN search engine's field search capability, we see a common usage pattern. Information specialists are the most frequent field searchers, with associate specialists only slightly behind. Administrators perform fielded searches much less frequently, while also using the IRN search engine less frequently on the whole (see Figure 6).

	All Catalog Fields	Title	Author	Doc #	Doc Type	Publisher	Servicename
Info. Specialists	3.8 (2-5)	4.7 (3-5)	3.0 (1-5)	3.3 (1-5)	3.8 (1-5)	4.0 (2-5)	3.7 (2-5)
Assoc. Specialists	3.1 (1-5)	3.6 (1-5)	2.9 (1-5)	2.9 (1-5)	2.7 (1-5)	3.3 (1-5)	2.4 (1-4)
Administrative	3.4 (1-5)	3.0 (1-5)	2.6 (1-5)	2.0 (1-5)	2.3 (1-4)	3.0 (1-5)	2.4 (1-4)
Average	3.6 (1-5)	4.0 (1-5)	2.9 (1-5)	2.9 (1-5)	3.2 (1-5)	3.6 (1-5)	3.1 (1-5)

Figure 16: Mean Rating of IRN Search Fields for Job, by Role and Field

(5-point scale, 1=Never, 2=Rarely, 3=Occasionally, 4=Frequently, 5=Must Have!)
Range of values is indicated in parentheses. "All Catalog Fields" is a conglomerate field of all metadata.

Specialists were asked to rank the value of available IRN search engine field options for their jobs. Information specialists gave the highest ratings for all fields, across the board (see Figure 15). They rated the title field as a required option, and showed a strong preference for the publisher, service name, document type, and "all catalog fields." By comparison, associates and administrators ranked the value of all fields significantly lower, across the board. For these users, no individual field averaged more than "occasionally" useful. These numbers echo the open-ended comments of staff members about the IRN search engine. Information specialists demand a wide variety of fielded searching options, while associates and administrators are more interested in a search engine that "makes sense" and delivers good results with minimal complexity. Clearly, any future interface design should take the needs of all these groups into account.

Staff members offered some additional guidance with their answers to the open-ended question, "What other searchable fields would be valuable for your work?" Information specialists mentioned additional support for two fields: abstract and date. Currently the search engine sorts by date, but does not explicitly offer date as a searchable field. Abstracts are available for some documents, but depend on manual additions by cataloguers. The abstract field should be a priority for future online cataloging. Administrators echoed information specialists, asking for keywords, abstracts, and date as searchable fields. Associates offered few comments, except for one who offered, “I normally don’t have to search for information in this way -- but I think it is a valuable tool for our information specialists.” This sentiment was reflected in the value ratings for individual fields. Overall, these ratings indicate that frequent searchers are largely satisfied with their field options, with the exception of a few specific fields like date, abstract, and keywords.

Implications

With my questions about frequency and value of fielded search options, I hoped to help establish the importance of metadata for IRN searchers. The results were promising, with a general consensus that the currently available fields are all at least occasionally valuable, with title and publisher the most frequently searched. Suggestions for additional searchable fields were realistic, and relatively simple to implement. Abstracts and keywords are already available for many documents; mainly this is an issue for the cataloguing team. Meanwhile date is already available as a sorting option, but is not an explicitly searchable field. It is a simple matter to create such a field. I will offer more conclusions about the importance of metadata fields, after we examine the search engine logs for concrete evidence of their utility.

Log Analysis of Fielded Searching

	All Catalog Fields	Title	Author	Doc #	Doc Type	Publisher	Service Name	All Searches
December 1998	68 (0.85%)	303 (3.79%)	6 (0%)	94 (1.18%)	325 (4.07%)	2945 (36.9%)	89 (1.11%)	7986
January 1999	167 (1.64%)	666 (6.53%)	26 (0.25%)	159 (1.56%)	384 (3.76%)	4943 (48.4%)	330 (3.23%)	10205
February 1999	215 (1.92%)	912 (8.13%)	31 (0.28%)	311 (2.77%)	452 (4.03%)	5159 (46.0%)	886 (7.90%)	11221
March 1999	267 (2.08%)	1026 (8.00%)	29 (0.23%)	355 (2.77%)	556 (4.33%)	1850 (14.4%)	1144 (8.92%)	12831
Total	717 (1.70%)	2907 (6.88%)	92 (0.22%)	919 (2.18%)	1717 (4.06%)	14897 (35.3%)	2449 (5.80%)	42243

Figure 17: Searches Performed per Field, by Field and Month
(Percentage of total is indicated in parentheses.)

After seeing the subjective value of the new metadata fields for IRN searchers, it is instructive to examine the log files stored by the search engine. Exactly how popular have the new field searching features been, since their introduction in December 1998? Figure 17 illustrates these figures, compiled from weekly log files for the last four months. Overall, some 56.1% of all IRN searches have involved at least one metadata field. The four most popular fields were Publisher (35.3% of all searches), Servicename (5.8%), Document Type (4.06%), and Title (6.88%).

	All Catalog Fields	Title	Author	Doc #	Doc Type	Publisher	Service Name
Info. Specialists	3.8 (2-5)	4.7 (3-5)	3.0 (1-5)	3.3 (1-5)	3.8 (1-5)	4.0 (2-5)	3.7 (2-5)
Assoc. Specialists	3.1 (1-5)	3.6 (1-5)	2.9 (1-5)	2.9 (1-5)	2.7 (1-5)	3.3 (1-5)	2.4 (1-4)
Administrative	3.4 (1-5)	3.0 (1-5)	2.6 (1-5)	2.0 (1-5)	2.3 (1-4)	3.0 (1-5)	2.4 (1-4)
Average	3.6 (1-5)	4.0 (1-5)	2.9 (1-5)	2.9 (1-5)	3.2 (1-5)	3.6 (1-5)	3.1 (1-5)

Figure 18: Mean Ratings of IRN Search Fields for Job, by Role and Field
(5-point scale, 1=Never, 2=Rarely, 3=Occasionally, 4=Frequently, 5=Must Have!)

These log figures correspond to the figures in Figure 18, where specialists selected Title and Publisher as the most valuable fields. Likewise, information specialists chose Document Type and Servicename as valuable fields, though the associates and administrators did not. The search logs revealed some surprising trends, as well. Though survey participants agreed on the value of an "All Catalog Fields" (conglomerate of all available metadata) search option, this field was rarely used (1.7% of searches). Document Number was only slightly more popular, with 2.18% of searches. However, this is more consistent with the searcher's rating of Document Number as an "occasionally" useful field. One anomaly deserves explanation: the Publisher field shrank from 46% of searches in February to 14.4% in March. This precipitous drop resulted from a new search interface, which was made specifically for materials from the IDC vendor. Previously IDC had been selected as the "vendor" on a generic market research interface; the new interface does not require this field, because it only searches IDC materials. Even after this change, Publisher remained the most popular search field.

Judging from the combination of search logs and survey results, it appears that the IRN's user community has gradually discovered and embraced the new metadata search options. With every available field, we see absolute search numbers which increase by the month, except for the Publisher field mentioned above. Servicename and Title are the most dramatic examples. Servicename only accounted for 1.11% of searches in December, but leapt to 8.92% by March, becoming the second most popular field. Title began with 3.79% of searches in December, and increased to 8.0% in March. Similar, but less dramatic, trends were seen with all the available metadata fields. In addition, we have seen an absolute increase in searches performed with the IRN search engine. The tool is becoming more popular, and a large number of searchers have begun taking advantage of its metadata search features.

Implications

For the most part, log analysis confirms the metadata preferences voiced by IRN staff. Among the entire user community throughout the corporation, several fields have dominated the searches: Title, Publisher, Servicename, and Document Type. The biggest surprise was the relative unpopularity of the "All Catalogue Fields" option. This may be the result of confusion about exactly what is searched by this option, which is meant as a conglomerate of *all metadata* available for a document. Most likely, usage of this field would increase with better labeling of its function. When we consider that 56.1% (23715) of searches during the past four months have involved metadata fields, it is obvious that these new options perform a valuable role for searchers of IRN online materials.

Ratings for Possible Future Features

	Topical Search Tree	Search Agents	Document Excerpts	Search Refinement	Query by Example	Multiple Field Sort
Info. Specialists	3.4 (2-5)	3.2 (1-5)	3.9 (3-5)	4.5 (4-5)	3.9 (2-5)	3.9 (1-5)
Assoc. Specialists	3.1 (2-4)	2.5 (2-3)	2.9 (1-4)	4.1 (3-5)	3.5 (2-5)	4.0 (3-5)
Administrative	4.1 (3-5)	3.0 (2-5)	3.3 (2-5)	3.7 (2-5)	3.7 (2-5)	3.7 (2-5)
Average	3.5 (2-5)	3.0 (1-5)	3.5 (1-5)	4.3 (2-5)	3.8 (2-5)	3.9 (1-5)

Figure 19: Mean Rating of Search Features for Job, by Role and Feature.

Range of values is indicated in parentheses.
(5-point scale, 1=Never, 2=Rarely, 3=Occasionally, 4=Frequently, 5=Must Have!)

When asked about possible features for the future of the IRN search engine, survey participants reiterated some of the preferences seen in their open-ended responses. Search refinement was the most popular option across all groups, seen as a "Must Have" by information specialists and highly desirable by others. Query by example, and multiple field sorting were also popular among all groups. Both of these ideas were mentioned by several participants in the open-ended responses. Personal search "agents" were the least popular feature among all groups, rating no better than "occasionally" useful. Two features showed mixed response between groups: topical search trees (like Yahoo), and document excerpts in result lists. Information specialists rated document excerpts "frequently" useful (3.9 with range from 3 to 5), while the other groups only saw them as "occasionally" useful. Most likely, this preference results from the large searches performed by information specialists, who must be able to evaluate the usefulness of search results very quickly. Likewise, a similar pattern is seen with the topical search tree feature. Administrators rated this feature "frequently" useful, while associate and information specialists ranked it "occasionally" useful. This would appear to reflect the different search styles employed by administrators, who are not so concerned about completeness of results. Whereas specialists tend to be seeking exhaustive information on a highly specific topic, administrators often want the *best* information on a more general topic.

	Market Research	Technical Standards	Regulatory /Government	Geographical	Corporate Internal
Info. Specialists	17 (94%)	7 (39%)	7 (39%)	6 (33%)	6 (33%)
Assoc. Specialists	7 (100%)	6 (86%)	1 (14%)	1 (14%)	3 (43%)
Administrative	7 (78%)	4 (44%)	0	0	4 (44%)
Total	31 (91%)	17 (50%)	8 (24%)	7 (21%)	13 (38%)

Figure 20: Information Sought for Job, by Role and Type
(Total percentage indicated in parentheses)

Features mean little without content, and survey participants gave strong indications of what sort of content needs to be covered by the IRN search engine. Two questions were asked about material coverage: "What types of information do you seek most frequently for your job?" (see Figure 17) and "What types of information need improved coverage by the IRN search engine?" (see Figure 18) The results help complete our picture of the IRN staff. Market research is searched frequently by all roles, for a total of 91% of searchers. Technical standards rank second, with 50% of total searchers. However, 86% of associate specialists search the technical standards, which is consistent with their role with the Call Center. Many of those calls are requests for printed copies of technical standards, and the associates are the "first line" for these questions. Corporate internal information was the next most popular search type, with 38% of total searchers, distributed evenly across all roles. Regulatory/government and geographical information were searched almost exclusively by information specialists, who frequently receive detailed requests for such information.

	Market Research	Technical Standards	Regulatory/Government	Geographical	Corporate Internal
Info. Specialists	12 (67%)	8 (44%)	9 (50%)	5 (28%)	9 (50%)
Assoc. Specialists	2 (29%)	3 (43%)	2 (29%)	1 (14%)	4 (57%)
Administrative	1 (11%)	2 (22%)	2 (22%)	2 (22%)	6 (67%)
Total	15 (44%)	13 (38%)	13 (38%)	8 (24%)	19 (56%)

Figure 21: Areas for Improved Search Coverage, by Role and Type
(Total percentage indicated in parentheses)

When asked about their preferences for areas of improved search coverage, IRN staff painted a different picture. All groups agreed, with 56% of the total vote, that corporate internal information needed improved coverage. Currently this type of information is handled by corporate AltaVista, and not by the IRN. This request reflects dissatisfaction with the official corporate search tool, and an eagerness for the IRN to take a new role. According to 50% of information specialists, regulatory/government information needed improved coverage, but this sentiment was not widely shared among other groups (29% of associates and 22% of administrators). A similar pattern is seen with market research: 67% of information specialists, with only 29% of associates and 11% of administrators. Both information specialists (44%) and associates (43%) agreed about the need for better coverage of technical standards, while only 22% of administrators felt so. Geographical information was the least controversial, with only 24% of respondents seeing it as an issue. Overall, we see a consistent desire for better coverage of corporate internal information, and of technical standards. Information specialists reinforced their status as a demanding user group, perceiving weakness in coverage of market research and regulatory information when others did not. These sentiments echo their open-ended responses, which demanded completeness of coverage for all online materials. It seems that information specialists will not be satisfied until every type of document, from every available vendor, is catalogued and indexed.

V. CONCLUSIONS

This study provided valuable information for the IRN, both as a user needs analysis and as an evaluation of current search tool effectiveness. We have identified three distinct groups of searchers within the IRN, who consistently demonstrate different information needs. Future iterations of the IRN search engine will need to take all these groups into account, with interfaces that are designed for maximum flexibility and clarity. Overall, the current IRN search engine registers as a significant improvement over its predecessor. However, much work remains to be done.

Throughout the survey analysis, a memorable lesson emerged involving search engine features vs. reliability. Again and again, with all types of search tools, we saw that the most exciting features do not translate to popularity for search tools. Most searchers prefer to stick with their “old reliable” search methods, because they *trust* certain tools to deliver good results consistently. Search engine features are secondary to this issue of performance. In the end, all searchers ask the same question, “Does it give me what I want, when I want it, consistently?” Judged by this standard, the IRN search engine still needs improvement.

What must we improve about the IRN search engine? Above all, searchers must be convinced that the IRN search engine is searching *all* available online materials. Though they like the improvements with fielded metadata searching, they want this metadata to be complete and consistent across all materials. When a searcher finds an important document with another tool, such as the IRN OPAC or the corporate AltaVista, they lose faith in the IRN search engine.

Therefore a top priority should be an ambitious cataloguing effort to standardize metadata throughout the collection, and to add metatags to all available documents. Once these goals have been achieved, searcher satisfaction should rise.

Metadata is not the answer for everyone. There are many good reasons why it has not gone very far on the Internet, mostly involving the difficulty of enforcing a common meaningful standard of description. However, a corporate digital library such as the IRN provides a superb environment for the implementation of metadata. With a trained cataloguing team and a large collection of documents on a fairly well-defined subject area (telecommunications), the IRN has the resources to create a superb archive of corporate information. Properly applied, metadata can offer a variety of access points to this information which are simply impossible with traditional full-text search tools. The IRN offers an excellent example of the promises and challenges offered by metadata. Digital libraries of the future will all contain some sort of electronic metadata, because it is a logical extension of cataloguing with online materials. Without catalogs, libraries could not exist. The IRN is taking the first steps in the direction of a truly digital library, and metadata promises to play an important part in this transformation.

VI. ACKNOWLEDGEMENTS

I could not have produced this study without the collaboration of dozens of my colleagues. I would like to thank the 38 survey participants from my corporate library, without whom this study would have been impossible. Thanks also to the several research specialists who volunteered to pre-test and critique my online survey. Their feedback was invaluable for the design of the survey. Likewise I must express my gratitude to Beverly Wiggins of the UNC-CH Institute for Research in Social Science, David Eckerman of the UNC-CH Institutional Review Board, and Evelyn Daniel of the UNC-CH School of Information and Library Science. Finally, I offer my hearty thanks to my advisor, Dr. Diane Sonnenwald, who patiently worked with me for more than a year on this project. Many times she offered her help on short notice, and I deeply appreciate her impressive generosity.

REFERENCES

Babbie, E. (1995). *The Practice of Social Research – 7th Edition*. Belmont, California: Wadsworth Publishing.

Drabenstott, K.M. (1996). Enhancing a new design for subject access to online catalogs. *Library Hi Tech, 14* (1), 87-109.

Fattahi, R. (1995). A comparison between the online catalog and the card catalog. *OCLS Systems and Services, 11* (3), 28-38.

Hakala, J. (1998, July). *The Nordic metadata project. Final report.* [Online]. Available: http://linnea.helsinki.fi/meta/nmfinal.htm.

Heery, R. (1998, March). What is … RDF? *Ariadne* [Online serial]. Available: http://www.ariadne.ac.uk./issue14/what-is/.

Infoseek Corporation. (1998). *FAQ: Ultraseek Server Usage* [Online]. Available: http://software.infoseek.com/products/ultraseek/faqs/faq059.htm.

International Federation of Library Associations and Institutions. (1999, January 11). *Digital libraries: Metadata resources* [Online]. Available: http://ifla.inist.fr/II/metadata.htm.

Lassila, O. & Swick, R. (Eds.) (1999, February 22). *Resource Description Framework (RDF) Model and Syntax Specification.* [Online]. Available: http://www.w3.org/TR/REC-rdf-syntax/.

Leighton, H.V., & Srivastava, J. (1997, June 16). *Precision among World Wide Web search services (search engines): AltaVista, Excite, Hotbot, Infoseek, Lycos* [Online]. Available: http://www.winona.msus.edu/library/webind2/webind2.htm.

Online Computer Library Center. (1997, October 2). *Dublin Core Metadata Element Set: Reference Description* [Online]. Available: http://purl.oclc.org/dc/about/element_set.htm.

Powell, A. (1997). Dublin core management. *Ariadne* [Online serial]. Available: http://www.ariadne.ac.uk/issue10/dublin/.

Powell, A. (1998, July). RDF and the Dublin Core. *UKOLOG, Manchester Conference Centre, July 1998.* [Online]. Available: http://www.ukoln.ac.uk/metadata/presentations/ukolug98/.

Schwartz, Candy. (1998). Web search engines. *Journal of the American Society for Information Science, 49* (11), 973-982.

UK Office for Library and Information Networking . (1999, March 10). *Metadata resources* [Online]. Available: http://www.ukoln.ac.uk/metadata/resources/.

Weibel, S., Godby, J., Miller, E., & Daniel, R. (1995, June). *OCLC/NCSA Metadata Workshop Report* [Online]. Available: http://purl.oclc.org/metadata/dublin_core_report.

World Wide Web Consortium. (1999, February 18). *Metadata and resource description* [Online]. Available: http://www.w3.org/Metadata/.

Younger, J.A. (1997). Resources description in the digital age. *Library Trends, 45* (3), 462-87.

Appendix A

Answers to Question: “What feature most distinguishes this product from other Internet search tools? Why might this feature be incorporated in the IRN search tool?

- (Yahoo) The directory structure is wonderful. Yes- I would love to have something like this for the IRN search tool- so that users could see some topical organization to our materials.
- (Metacrawler) Access speed
- (Northern Light) Categorises results into folders (eg commercial sites- educational sites etc) and allows accurate search terms. I find it often brings up the same sort of articles I find in Dialog or RBB (although not always for free).
- (Hotbot) HotBot often will give the top 10 most visited sites for a given search string. These are typically right on target. This feature could help target the most relevant hits within the IRN based on similar queries.
- (AltaVista) Reliability and comprehensiveness
- (Northern Light) Folder tree view of results- makes large of returns maneagable- leads to more natural search refining. Good results for ad-hoc plain english queries. Folders would be a good way to categorize topics- point people to related sources/types of information.
- (Yahoo) I guess I am used to it
- (Yahoo) Can find URL of any organisations easily
- (Metacrawler) searches all the other search engines I don't do a lot of detailed searching on the web. Most often the information I am looking for is not free. There is nothing from this search tool that I would like to see incorporated into the IRN search tool.
- (AltaVista, Hotbot, Google) AV frequently gives me results others miss. Boolean searching is also easy. Don't get as many inexplicable hits as I do with other search engines. Fast. However- I prefer lots of HotBot's features--the display of results is easier to understand and the ability to specify dates or media types is helpful. It's also easy to modify/narrow a search with HotBot and HotBot lets you see long scrolling lists of results instead of 10 or 20 hits at a time. (HATE that page by page deal in other products.) Google is another favorite. Haven't used it much yet (not sure how extensive the coverage is and haven't had time to investigate) but searching is fast and intuitive and relevancy seems consistently superior to the other two. For example- put in a company name with google and I get the company web page as the first hit. Usually- that's what I want to see first. Category searching and field sorting options would be good for the IRN engine. Sorting by date is most critical and by author/publisher next for me. Also need
- (Excite) The 'more like this' feature is very useful.
- (AltaVista) Would not want as IRN search tool
- (Infoseek) I like Infoseek because the page I am looking for usually appears within the first 5 listed results.
- (AltaVista) I can usually find standards information required or Standards Bodies websites and other links which are helpful to have. Many different languages and translations are available for many of these sites.
- (Yahoo) Durable URL's for company financials- etc. Integrated with other things I use a lot like stock watch
- (Hotbot) Seems to be able to handle more complicated search statements.

Answers to Question: "What feature most distinguishes this product from other commercial search tools? Why might this feature be incorporated in the IRN search tool?

- (Dow Jones) Excellent coverage of late-breaking news- with good sorting of results. IRN needs more up-to-date materials coverage.
- (Dialog) cross file searching across all selected sources. The IRN search engine still only searches some market research- some newsletters- not the catalogue- not any newsfeeds- not Onesource or telegeorgraphy- not standards. etc
- (Dialog) key words in context field
- (Reuters) Its pre-set categories- ease of use and presentaion of results make it a very good news source. It's not very good for complex Boolean searches which I would use Dialog for.
- (IDC-Net) By clicking on any of the indexed terms in one document- one can immediately pull up other documents with those same indexed terms. It really helps to maneuver through related documents- especially when they would not have popped up in the original search. It also allows searches by module and gives a browisng feature by module. Both of the above features could be incorporated in the IRN search engine to better pinpoint results that are highly pertinent.
- (Dialog) Comprehensive content and varied search options and ability to search multiple sources simultaneously
- (Dow Jones) interface friendly use; could narrow down the search
- (Dialog) Complexity of the search strategies that can be created impacts retrieval. We need to have better search syntax in our search engine for searching full-text documents.
- (Dialog) Best feature is the ability to create and maintain sets; then mix them for new results sets- ie. 'building block searching.' For example- I create a set around a concept; then additional sets around additional concepts; then match them up in different ways- generally paring hits down to an acceptable number with good precision. Next best feature is that Dialog searches across as many as several hundred databases at once with a single strategy. I generally use about 10 databases on most searches. This is very helpful because I can eliminate duplicates in my final set. If I ran the search 10 times in 10 databases- I'd have lots of copies of the same articles. The subject coverage in Dialog is the broadest of any vendor's. Another feature is the ability to create Reports and Rankings. You can rank your search results to give you largest companies first- for example- creating a report that can go directly to a client. Dialog also has an excellent help desk with specialists who will- for free- help you create sea
- (Dialog) Availability of a host of searching capabilities.
- (Dialog) I use Dialog more often than other commercial databases because I find the content most relevant to what I am looking for. Dialog features that I like that could be incorporated in IRN search tools are: flexibility in using search operators and proximity operators- ability to search by field- ability to combine and reuse search sets- segmenting out parts of sets by date and keywords- and different databases for different types of information.
- (Dialog) Dialog is very accurate due to its excellent indexing.
- (IHS Standards)This is a standards full text tool which allows us to download standards we have subscribed to through an EXTRANET LINK.
- (Dow Jones)Best source for near real time awarensess. Events usually cause my inquiry.

Answer to Question: What feature(s) most distinguish the IRN Web Site Search Engine from similar site search tools?

- Excellent category search support. I like searching on document number- and publisher.
- lack of searching everything on the site. It cannot claim to be a site search tool as it is looking at limited areas. It is misleading at the head of the site and should only appear when in the features it searches.
- pull-down menus make the search engine easier to use
- i don't use excalibur very much and don't have much experience with the tool. i don't think my answers should be taken into consideration.
- After using Northern Light/Alta vista I find Excalibur slightly more imprecise (eg I'm not sure of the difference in results between phrase searching within " and without")- but I think this is perhaps more to do with the cataloguing. My main grumble is the way the Reports/Bulletins button keeps reverting to 'All Types' after I do an initial search (I hate bulletins!). That said- I like the Vendor/Sorting/Style options- although I'm never quite sure how precise the catalogue fields are (eg does telecom=telecommunication?)
- It gives the option of searching by vendor and includes Meta tags.
- The category searching and sorting by vendor I also like the result box illustrating the key words
- Precice responses. Not necessarily complete responses...
- Nothing I can think of
- It retrieves items with very low relevance all to frequently. Have little understanding or confidence about how or sometimes even what it is searching.
- Results make sense. My only concern is that I'm getting complete results. Great improvements have been made but I'm still not confident. I tend to stick to the most basic strategies I can. The more features I use- the less I trust the results. Also- I need a better understanding of what market research is and isn't covered when I do a search.
- Not happy with results in general--not very relevant.
- I don't really use this search engine since it only has marketing type of information and I don't require to do any work within that area.
- Searches market research?
- Has improved greatly.

How would you improve the IRN Web Site Search Engine? What additional capabilities would you add?

- I would prefer to see improved searching of PDF documents.
- I would look for a new search engine. Let's not hold on to Excalibur just becausew it is available to us.
- Searching across all feeds/features on the site or at least better inclusion/exclusion notices seen by all users- not just advanced
- I did like the 'refine' feature which is no longer available (except according to the help menu).
- I'd have it recognize truncation on its own. In other words- if I put in the term CLEC- it would pull CLEC and CLECs without me needing to specifiy a wildacrd at the end. I'd also have it recognize any style of searching- without needing to change the style. So if I typed in a phrase- it would recognize that string as a phrase or if I wanted to use Boolean- it would know it was Boolean. It would also be nice to find related documents of the one I find really useful - like IDC's website tool.
- You could have an advanced search button which would have several entry boxes separated by adjustable boolean operators--this would assist the end users. I still need to increase my confidence in the search engine; the improvements of late are fantastic...it's just difficult to let go of the uncertainty-'did it pull up everything'.
- More options on the main search page. 'What's related' or 'Browse Sections' on results page.
- 1- better grouping of search results. 2- more tips for searching. 3- search engine should never have any pop-ups asking for securtiy verfication. This should be done in the lower level 4- more bolean type searching
- I really don't use it that much for detail work .It's fine for my needs.
- The 'help' manual should explain its search facilities more clearly with examples; adding more search capabilities to narrow down a search.
- Could we have things like in the Dow Jones search engine- percent or weighted relevance. In most cases you don't want to look at records with one reference to Sonet when you are looking for info on Sonet. I find the engine that Dow Jones uses to be one of the best of all the different types of engines I search.
- Education would help. I do not believe I am an expert. Don't have a full understanding of the coverage and capabilities. Otherwise- see above.
- Field searching.
- Ability to search both by date and relevancy at same time - say- ability to eliminate items published before 1998. Better relevancy ranking - would prefer items that have the keywords in the title to appear at the top of the list- before older items that have the keyword buried in the text. Personally would prefer large market reports to appear at top of list compared to financial times newsletters. More ability to search by field and use search and proximity operators - more like Dialog or the IRN catalogue.
- Search strategies should be similar to other search engines - excalibur is too 'idiosyncratic'. Need to feel confident that engine is searching everything.
- It would be nice to have it actually tell you what that feature does search and what type of documents it searches. Tips on how that searches the IRN website and what capabilities are available. I didn't find out until a week ago what that search engine actually searches. I found out it only searches marketing type of documents. Is that a correct assumption? I know from clients feedback calling me and stating they did a search within the IRN Website and could not find standards information not realizing that it does not even search that type of documentation. I would like to know why that feature is even there if it does not search anything else but marketing documentation. What capabilities are available? It would be nice to be able to search for example standards documentation. EXAMPLE: IEEE 802.3 STANDARD search by 'keyword' or 'Document number' TO RETRIEVE RESULTS and refer clients to 'Infocat' if document available within the IRN and/or refer clients to 'IHS WWSS' for full text downloadable s
- Combine with a 'Yahoo-like' content tree. Better response to plain language 'need...' type questions

Appendix B

Examples of IRN Search Engine Log Entries

Entry for 12/30/98 search of IDC publisher, for term “monitoring” in title field. 2 documents returned.

```
R: 621
U: "IDC"
T: 12/30/98 13:11:48.778 -- 13:11:49.116 (0.338)
A: SQRY
Q: title= monitoring
| dc_publisher= "IDC"
| dc_relationtype= IsParent
L: "idc" (DOCS=2, QW=3, T=0.317)
| "mrr_meta" (DOCS=2, QW=3, T=0.317)
| "datapro_meta" (DOCS=2, QW=3, T=0.317)
| "forrester_meta" (DOCS=2, QW=3, T=0.317)
| "ft_meta" (DOCS=2, QW=3, T=0.317)
I: 18429
C: 2/30000
```

Entry for 12/30/98 search of IDC publiser, for document number 17309. No documents returned.

```
R: 625
U: "IDC"
T: 12/30/98 13:12:21.614 -- 13:12:21.958 (0.344)
A: SQRY
Q: dc_identifier= 17309
| dc_publisher= "IDC"
| dc_relationtype= IsParent
L: "idc" (DOCS=0, QW=2, T=0.326)
| "mrr_meta" (DOCS=0, QW=2, T=0.326)
| "datapro_meta" (DOCS=0, QW=2, T=0.326)
| "forrester_meta" (DOCS=0, QW=2, T=0.326)
| "ft_meta" (DOCS=0, QW=2, T=0.326)
I: 18431
C: 0/30000
```

Entry for 4/12/99 search for “Parallel Computing” in all catalogue fields (rdf_metadata). 386 documents returned.

```
R: 71
U: "GUEST"
T: 1999-04-11 14:33:15.861 -- 14:33:15.281 (2.419)
A: SQRY
Q: rdf_metadata= Parallel Computing
| dc_relationtype= IsParent
L: "datapro_meta" (DOCS=386, QW=11, T=1.113)
| "forrester" (DOCS=386, QW=11, T=1.113)
| "idc_meta" (DOCS=386, QW=11, T=1.113)
| "mrr_meta" (DOCS=386, QW=11, T=1.113)
| "ft_meta" (DOCS=386, QW=11, T=1.113)
I: 12341
C: 386/30000
```

Appendix C