Information Resources: Modelling, Cataloging and Searching*

Bipin C. Desai†, Yves SAINTILLAN, Rajjan SHINGHAL
Department of Computer Science,
Concordia University,
1455 de Maisonneuve Blvd. West,
Montreal, H3G 1M8 CANADA
Email:{bcdesai,y_saint,shinghal}@cs.concordia.ca

June 1995

Abstract

Existing search systems exhibit uneven selectivity when used in seaking information resources on the Internet. This problem has prompted a number of researchers to turn their attention to the development and implementation of matadata models for use in indexing and searching on the WWW and Internet. In this paper, we present our results of a simple query on a number of existing search systems and then discuss a proposed metadata structure. Modelling the expertise of librarians for cataloging, user entry and search using a rule-based system is also discussed.

1 Introduction

Access to relevant information is one of the most important requirements of all human activities. This need has been recognized and has resulted in the continuing effort to catalog and organize information so as to facilitate it's expected discovery and access. An increasing number of research institutes, universities and business organizations are currently providing their reports, articles, catalogs and other information resources on the Internet in general and on the Web [BERN, BERN2] in particular. This is now becoming the accepted method of disseminating and sharing information resources in hypermedia. At this time a number

*Address all correspondence to the first author
†URL: http://www.cs.concordia.ca/~faculty/bcdesai/

of information sources, both public (free) and private (available for a fee), are available on the Internet. They include text, computer programs, books, electronic journals, newspapers, organizational local and national directories of various types, sound and voice recordings, images, video clips and scientific data. Also, private information services such as price lists and quotations, databases of products and services, and specialty newsletters are available.

A number of index generation systems and related search systems are currently available on the Internet [DACL, DEBR, EMTA, FLET, GNAM, HARV, INFO, KAHL, KOST, MAUL, MCBR, NIKO, RBSE, THAU, WEBC, WWWC, YAHO]. Some of these are manually generated indices(Aliweb [KOST], CUI W3 Catalog [WWWC], GNA Meta-Library [GNAM], DA-CLOD [DACL]), while others are generated by robots (Harvest [HARV], InfoSeek [INFO], Lycos [MAUL], Nikos [NIKO], RBSE [RBSE], Web Crawler [WEBC], WWWW [MCBR], Yahoo [YAHO]). Some of these robot based system (e.g., WWWW [MCBR]) also allow manual entries. Some of these are specialized for the Web, others are for locating files on Anonymous FTP sites. The search interface provides users very little flexibility and the results obtained are varied. This is illustrated in Table 1 for a query using the first and last names of the author as the search term. Even Lycos which is reported to have indexed nearly 4 million documents has had only partial success in locating all relevant documents [1].

2 The Problem

The purpose of indices and bibliographies (called secondary information) is to inventory the primary information and to allow easy access to it. The traditional method of generating bibliography entries required finding the primary source, identifying it as to its subject, etc., describing it for later matching for unknown future users, and classifying it according to accepted norms.

The unpredictable retrieval of appropriate information resources documented in Table 1 illustrated that there is a need for the definition of suitable matadata to model resources and associated support systems which allows better controlled 'search for' and 'access to' these resources. With the current plethora of index services and search systems, most users are lost. Even after a search there is no guarantee that the appropriate information resource will be found. Furthermore, these systems are not able to function together due to the differences in their coverage, indexing structures and user interfaces.

A number of projects in the Library domain have addressed the problem of cataloging and in particular the cataloging of information in electronic and multi-media format. CORE [CROM], MARC system [BRYN, CRAW, MARC, PETE], MLC[HORN, ROSS, RHEE] and TEI [GAYN, GIOR] are examples of some of these initiatives. These existing and proposed indexing systems range from a minimum to a full level of bibliographic information. However, such systems are designed for professional catalogers and many of the elements included in

[1] The list of URLs of document known to contain the string is given in [DESA3]. The tests for all systems except InfoSeek and WWWW were done on June 3, 1995; the tests for InfoSeek and WWWW were done on June 15, 1995. Result may not be identical if the tests were to be repeated due to the possible discovery of the missing documents by the index systems involved. **All** documents in [DESA3] existed *well* before the test dates.

Search System	**Number of Hits**	**Number of Duplicates**	**Number of Mis-hits**	**Number missed**
Aliweb	none	-	-	24
DA-CLOD	none	-	-	24
EINet	6	0	4	22
GNA Meta Lib.	none	-	-	24
Harvest	none	-	-	24
InfoSeek	7	0	0	17
Lycos	231	2	222	17
Nikos	none	-	-	24
RBSE	8	-	8	24
W3 Catalog	none	-	-	24
WebCrawler	7	3	0	20
WWWW	2	0	0	22
Yahoo	none	-	-	24

Table 1: Search statistics for using search the term Bipin (AND) Desai *NOTE: The misses in the results for the manual systems, some of which depend on registering the resources, indicate that the resources have not been registered.*

them, though useful, are beyond the comprehension of most providers or users of information.

3 Semantic Header

Professional catalogers have found the need for elements similar to those in [DESA2] in most indexing applications. This dictates that they must be included in most indexes for information resources. The dependence on titles as the most common search criteria dictates that they must be indicative of the contents of the document. This is not always the case, hence someone (the author or the cataloger) has to include annotations, keywords or key phrases to indicate the actual content.

Accuracy or quality of a document can be indicated by including reviewers' opinions. Such opinions, however, are rarely accessible to the traditional cataloger. Another feature of importance to the user of an index is the presence of an accurate abstract. An abstract provides a summary of the material and thus is more indicative of the contents than the title or keywords supplied by the author, bibliographer or that selected from scanning the text. Reference librarians and library users tend to use such annotated bibliographies in selecting from competing sources. Fortunately, for an on-line index system as proposed in CINDI [DESA1], it is possible to include not only the author supplied abstract but also annotations made by independent users in the index entry for the information resource.

Semantic Header [DESA] was conceived as a required component of all HTML documents for the Web. It was originally presented at the First International World Wide Web Conference in Geneva in April 1994. Since then, it has been extended to resources accessible directly on the Internet.

The structure of the index is similar to the ones used for most library indices and include other extrinsic information deemed useful for on-line systems. The syntax of the semantic header is similar to the HTML markup language [BERN1] which is based on the SGML [GOLD] markup language. However, the user working with the index entry system is guided through the process by an expert system. This system guides the user in the choice of standardized terms through an easy to use graphical interface.

The Document Type Definition (DTD) uses a set of rules in the description of the document content. Some of these are listed below:

- An element followed by the symbol "?" is optional and could occur zero or one time.
- An element followed by the symbol "*" can occur zero or more times. This symbol is called the *optional and repeatable occurrence indicator* [GOLD].
- An element followed by the symbol "+" is required and can occur one or more times. This symbol is called the *required and repeatable occurrence indicator* [GOLD].
- The percent sign is used to define a subset of elements before its first reference. Example:

  "%ROLE" is the reference to the entity named ROLE. Each entity should be declared between tags as follows:

```
<!ENTITY % ROLE  "Author | Artist | Co-author | Corporate Entity |
                  Editor |Designer |Programmer | Publisher | Other">
```

- Data types and keywords:

  #PCDATA : zero or more parsed data characters

  CDATA : character data

  EMPTY : empty content

  The DTD for the Semantic Header is given in Figure 1 below.

3.1 Semantic Header: Advantages

The proposed Semantic Header and the discovery system based on it exhibit the following advantages.

- The Semantic-Header allows the indexing of resources accessible by any protocol.

```
<!-- Parameterizable list -->
<!ENTITY % ROLE  "Author | Co-author | Editor | Artist | Corporate Entity | Designer |
                  Programmer | Publisher | Other">
<!ENTITY % ID  "FTP | ISBN | ISSN | Gopher | HTTP | SHN | UAS | URN | Other">
<!ENTITY % CLASSIFICATION  "Legal | Security | Nature | Other">
<!ENTITY % COVERAGE  "Audience | Geographical Coverage | Spatial Coverage | Epoch | Other">
<!ENTITY % SYSREQ  "Hardware | Network | Software | Other">
<!-- Element list -->
<!ELEMENT Semhdr --
         (Title, Alt-title?, Subject+, Language?, Char-Set?, Author+, Keyword+, Dates,
          Version?, Supersede?, Classification*, Coverage*, Sysreq*, Genre?, Source?,
          Cost?, Abstract?, Annotation?......>
<!ELEMENT Title           --          (#PCDATA) >
<!ELEMENT Alt-title       --          (#PCDATA) >
<!ELEMENT Subject         -0          (EMPTY)   >
<!ELEMENT Language        --          (#PCDATA) >
<!ELEMENT Char-Set        --          (#PCDATA) >
<!ELEMENT Author          -0          (EMPTY)   >
<!ELEMENT Keyword         --          (EMPTY)   >
<!ELEMENT Identifier      --          (#PCDATA) >
<!ELEMENT Dates           --          (Created, Expiry?, Updated?) >
<!ELEMENT Created         --          (#PCDATA) >
<!ELEMENT Expiry          --          (#PCDATA) >
<!ELEMENT Updated         --          (#PCDATA) >
<!ELEMENT Version         --          (#PCDATA) >
<!ELEMENT Supersede       --          (#PCDATA) >
<!ELEMENT Classification  --          (#PCDATA) >
<!ELEMENT Coverage        --          (#PCDATA) >
<!ELEMENT Sysreq          --          (#PCDATA) >
<!ELEMENT Genre           --          (#PCDATA) >
<!ELEMENT Source          --          (#PCDATA) >
<!ELEMENT Abstract        --          (#PCDATA) >
<!ELEMENT Annotation      --          (#PCDATA) >
<!ELEMENT Cost            --          (#PCDATA) >
<!ELEMENT Control         --          (#PCDATA) >
 <!-- Attribute List -->
<!ATTLIST Subject        General            CDATA             #REQUIRED >
                         Sublevel1          CDATA             #IMPLIED  >
                         Sublevel2          CDATA             #IMPLIED  >
<!ATTLIST Author         Role               (%ROLE;)          #REQUIRED >
                         Name               CDATA             #REQUIRED >
                         Organization       CDATA             #IMPLIED  >
                         Address            CDATA             #IMPLIED  >
                         Phone              CDATA             #IMPLIED  >
                         Fax                CDATA             #IMPLIED  >
                         EMail              CDATA             #IMPLIED  >
<!ATTLIST Identifier     Domain             (%ID;)            #REQUIRED >
<!ATTLIST Classification Domain             (%CLASSIFICATION;)#REQUIRED >
<!ATTLIST Coverage       Domain             (%COVERAGE;)      #REQUIRED >
<!ATTLIST Sysreq         Component          (%SYSREQ;)        #REQUIRED >
<!ATTLIST Annotation     Signature          CDATA             #REQUIRED >
```

Figure 1: DTD for Semantic Header

- The Semantic-Header will include annotations of reviewers and other users thus providing the possibility of a better informed decision regarding the source resource.
- The Semantic-Header syntax offers a way to register standardized keywords chosen by the provider of the resource. The existing search systems often hack terms from title and/or content resulting in unpredictable results.
- The Semantic-Header has an "abstract" element which provides a better idea of the resource than an excerpt prepared by systems such as Lycos.
- The number of documents stored is not limited since the database is distributed amongst different sites.
- The Semantic-Header is designed to form part of each resource. The HTML/SGML based syntax allows its display by appropriate browsers.
- In existing indexing systems, one of the limitations is the low number of indexed documents. This is illustrated by the disappointing results for the manual index systems such as ALIWEB in Table 1. The difficulty lies in convincing people to register information regarding their resources. This problem would be solved in the Cindi system by:
  - providing an easy-to-use interface to register matadata for resources, and
  - assuring the presence of matadata in resources by browsers.
- Since the registration of the Semantic-Header in the database is performed by the provider of the resource, it has the following advantages:
  - costwise
  - accuracy
  - efficiency

4 Cataloging and Searching: Modelling the Expertise of a Librarian

The expert advice offered by a librarian should be offered to a user of our application for both searching and cataloging. In searching for a given set of documents, often the user offers vague, partial and/or incorrect information in her/his attempt to identify the terms used for the various descriptors of the index for the documents for which s/he is searching. In other words, the user search specification is often "ill-structured"; hence, expertise is needed to help users to articulate their needs. In cataloging a new resource, the cataloging librarian uses knowledge of authority and accepted norms for classification. From such knowledge s/he chooses terms to describe the resource. Reference librarians are aware of the conventions used by catalogers as well. They are typically familiar with the classification schemes, terms, index structures, and resources available in the domain of the user's need. The expertise of a reference librarian should be replicated to assist the users of our application in both searching and cataloging.

To model the expertise of a reference librarian we use a small embedded expert system in our current implementation. Expert systems have been used quite extensively whenever human intelligence needed to be modelled. The intelligence of a reference librarian used to identify and process a user's need for authoritative terms for lay terms (for both searching and cataloging) can be represented as a set of "`if...then...`" rules called the rule base. The rule base thus comprises the acquired expert knowledge [CHAN,GIAR,LUCA,SHIN].

In the current application on cataloging, search, and discovery of documents over the emerging international information infrastructure, the expert system assists the user in entering the controlled terms by providing a context-sensitive help mechanism in the user interface, and also helps to formulate a search query. We refer to the two aspects of the expert system as the User Interface (UI) part and the Cataloging/Search Expertise part respectively. Each is meant to capture a different perspective of the mental view held by a librarian. Owing to the space limitation, we focus our discussions solely on the approach we use to provide expertise for discovery and search.

The user interface for searching for resources in the information infrastructure in our application requires the user to input information that will help identify semantic headers of resources distributed over the network. The user entry can be at different levels of detail, and depending upon the level of detail entered, the UI part of the expert system provides the required amount of help to complete the input. This amounts to a reference librarian guiding a novice user in entering data based on the current level of data that has already been entered by the user. For example, if the user has entered the subject `Computer Science`, then the help for other fields would be tailored for `Computer Science`. Later, if the user changes the subject to `Chemistry`, help information would change accordingly.

The context sensitivity of help is complicated because the user can input a *synonym* for an input field, a case in point being "subject". In addition, synonyms can be entered for any of the fields corresponding to the subject hierarchy. A synonym must be resolved in an appropriate **control vocabulary** so that it would make a match with an appropriate field on the semantic headers of documents. For example, the user can enter the synonym `KBS` which can mean `Knowledge Base Systems, Expert Systems`, or `Deductive Data Base Systems` that are part of the control vocabulary. In general, searching the subject hierarchy for a control vocabulary can be modelled as a graph (tree) traversal. Synonyms, do however, complicate this search for a control vocabulary term because they would make this tree traversal into a directed acyclic graph traversal which is computationally demanding. Yet synonym resolution allows the system to automatically fill in higher levels of the subject hierarchy as they can be determined uniquely once a synonym at a lower level is resolved in a control vocabulary. This greatly aids focusing the search to a specific set of documents. For example, let `KBS` entered for level-1, associated with the subject hierarchy, be resolved to `Expert Systems`, so that the subject of the entry would be automatically updated to contain `Artifical Intelligence` under which `Expert Systems` is catalogued. In order to facilitate resolution of synonyms, the expert system displays a list of control vocabulary items at any level in the subject hierarchy whenever a synonym is entered. The synonyms and the associated control vocabulary are kept in a local database.

A second source of complication arises when the user enters *partial values*: a substring for subject, for example `Data Bas`. Though one can display a list of subjects that have this substring, context sensitivity implies that the partial values already entered in other fields must also be considered in providing a help response to the user. Thus, the help would be based on not only the partial values of the current field, but also on existing values

of other related fields. For example, if the user has entered `Hybrid relational` in level-1 and `Frame` in level-2 field, then the context sensitive help for subject would take into account the current values in level-1 and level-2 fields before providing appropriate help to the user. Had the current values at level-1 and level-2 been ignored, then the user would be provided with a long list of subjects, many of them would have level-1 and/or level-2 of their hierarchy that would not match the current values of the level-1 and level-2 fields. This is equivalent to capturing a reference librarians mental view to help focus the search for documents. The values in the other related fields need not be full values, but can themselves be partial values. Context sensitivity in the UI part of the expert system gives only the appropriate amount of help that would be needed at a particular point in time.

The third aspect of the expertise is *automatic inferencing capability*. A reference librarian would be able to identify that the user is searching under the subject `AI` if the user tells the librarian that s/he is searching for documents about `expert systems`. In the system, once the value entered for a particular field is complete, but the other related fields are empty, the pertinent values of these fields should be inferred and automatically updated with those appropriate entries. Such automatic inferencing of other related fields would help to not only focus the search query to select a smaller number of documents, but also in query optimization.

The final aspect is *warnings*. Whenever a user-entered field does not match the existing field values, the user should be warned because this could result in a search query whose processing would not produce any document retrieval. For example, `Expert Systems` as a level-1 entry does not match with `Chemistry` as a general subject (level-0), and the user must thus be warned of this potential mismatch. This doesn't mean that specific expert systems cannot exist in Chemistry; all it means is that expert systems is not a sub-subject of chemistry. This mimics the cautionary advice that a reference librarian would give to a user.

We have developed a small demonstration version of the UI part of the Expert System for rule testing, refinement, and for eventually evolving it into a complete system. The access system is simulated using a table and a set of functions. The expert system is coded under Motif; the rules themselves are implemented in C. A sample rule that handles the context sensitivity of a partial value entered for subject is shown in Figure 2. The rule is meant for checking partial sub-strings entered by the user for the subject field, and based on the current partial values of the level-1 and level-2 entries, an appropriate list of subjects are displayed from which the user may choose. Other rules in the system take care of the other aspects of the expertise: synonym resolution, automatic inferencing, and warnings. For example, if the user enters `Chess` for level-2 of the subject hierarchy, then level-1 and level-0 entries would be automatically filled with `Game playing`, and `Artifical Intelligence` respectively. Similarly, whenever a user-entered field of the subject hierarchy does not match with the current, possibly partial, values of other fields, an appropriate warning is issued.

The query formulation, that is the search part of the expertise modelling, would formulate a search query based on the current value of the fields entered. The result of the query would be a set of semantic headers, possibly empty, matching the user search request. Before formulating a search query, however, additional checking is made: this is typical of the way a reference librarian would proceed to focus his search to identify a smaller number of documents: for example, ask the user to enter author information, title information, spelling correction, phonetic checks, and input checks such as mismatched subject hierarchy, etc. If the result of the search query is 0 semantic headers, then this is displayed as "No documents

```
void Rule2(char *subject_string)   /* context sensitive subject help */{

    Widget list_dialog;
    char **subject_list;

    char *l1 = NULL, *l2 =NULL;

    /* rule 2: if not synonym, check list of subjects matching this
       sub-string */
    if (get_synonym(subject_string, 0))
      return;

    l1 = XmTextGetString(level_1);
    l2 = XmTextGetString(level_2);

    /* context sensitivity: get only the matching list */
    subject_list =  get_matching_subjects(subject_string,l1, l2);
    if (subject_list[0])
    {
         list_dialog = create_list_dialog(0);
         add_items(subject_list, list_dialog);
    }
}
```

Figure 2: A sample rule encoding in the test system for context sensitive help associated with the general subject field entry in the user interface. Similar rules code other context sensitive help such as synonym resolution, automatic inferencing, and warnings.

found"; otherwise, the user is allowed to pick a subset of the retrieved documents for display.

5 Conclusion

Using a typical query as an example, we have presented statistics on a number of known Interenet search systems. To improve the efficacy of search, we have developed a semantic header: a data structure to record the matadata of network resources. Provided by the author/creator of the resource, it not only indicates the contents of the document and provides some vital extrinsic attributes, but also helps in indexing and locating the document. To aid a typical user's search for resources on the Web, we have developed an expert system modelling the expertise of a reference librarian. The expert system guides users, who may have only incomplete information on the resources they are trying to locate, in formulating

their queries to locate a set of semantic headers and thence the relevant resources.

6 Acknowledgments

An initial version of the expert system was developed by Dao Nguyen. It was later updated and revised by P. Gokul Chander, who developed the current version of the expert system described in section 4. The knowledge base of the expert system was developed after consulting Carol Coughlin and Lee Harris, reference librarians at Concordia University. The help from these people is gratefuly acknowledged. This work is supported, in part, by a grant from Seagram's Fund for Academic Innovation.

7 References

[BERN] Berners-Lee, T., Cailliau, R., "WorldWideWeb: Proposal for a HyperText Project", http://info.cern.ch/hypertext/WWW/Proposal.html

[BERN1] Berners-Lee, Tim, Connolly, "Hypertext Markup Language, Internet working draft", http://info.cern.ch/hypertext/WWW/MarkUp/HTML.html

[BERN2] Berners-Lee, T. "Wide Web Initiative: The Project", http://info.cern.ch/hypertext/WWW/TheProject

[BYRN] Byrne, Deborah J., "MARC manual: understanding and using MARC record", Libraries Unlimited, Englewood, Colo. 1991.

[CHAN] Chander P. G, Shinghal R., and Radhakrishnan, T. "Goal Supported Knowledge Base Restructuring for Verification of Rule Bases", In Notes of the Workshop on Validation of Knowledge-Based Systems (Fourteenth International Joint Conference on Artificial Intelligence), Montreal, Canada, August 1993. In press.

[CRAW] Crawford, Walt, "MARC for Library Use: Understanding USMARC", G. K. Hall, Boston, MA, 1989.

[CROM] Cromwell, Willy, "The Core Record: A New Bibliographic Standard", Library Resources and Technical Services, Vol. 38-4, pp. 415-424, 1994.

[DACL] DA-CLOD, http://schiller.wustl.edu/DACLOD/daclod

[DEBR] De Bra, P., Houben, G-J., & Kornatzky, Y., "Search in the World-Wide Web", http://www.win.tue.nl/help/doc/demo.ps

[DESA] Desai, Bipin C., "Cover page aka Semantic Header", http://www.cs.concordia.ca/semantic-header.html (July 1994), http://www.cs.concordia.ca/~faculty/bcdesai/semantic- header.html (revised version, August 1994)

[DESA1] Desai, Bipin C., "The Semantic Header and Indexing and Searching on the Inter-

net", February 1995, http://www.cs.concordia.ca/∼faculty/bcdesai/cindi-system-1.0.html

[DESA2] Desai, Bipin C., "Report of the Metadata Workshop, Dublin, OH", http://www.cs.concordia.ca/∼faculty/bcdesai/metadata-workshop-report.html, March 1995.

[DESA3]Desai, Bipin C., "Test of Internet Indexing Systems", http://www.cs.concordia.ca/∼faculty/bcdesai/www/test-of-index-systems.html

[EMTA] Emtage, A., Deutsch, P., "Archie: An electronic directory service for the Internet", Proc. Winter 1992 Usenix Conf., pp 93-110, 1992.

[FLET] Fletcher, J. 1993., "Jumpstation", http://www.stir.ac.uk/jsbin/js

[GAYN] Gaynor, Edward, "Cataloging Electronic Texts: The University of Virginia Library Experience", Library Resources and Technical Services, Vol. 38-4, pp. 403-413, 1994.

[GIAR] Giarratano, J., and Riley, G. "Expert Systems: Principles and Programming (2nd edition)", PWS Publishing Company, Boston, MA, 1993.

[GIOR] Giordano, Richard, "The Documentation of Electronic Texts Using Text Encoding Initiative Headers: An Introduction", Library Resources and Technical Services, Vol. 38-4, pp. 389-401, 1994.

[GNAM] Global Network Academy Meta-Library, http://uu-gna.mit.edu:8001/cgi-bin/meta

[GOLD] Goldfarb, Charles F., The SGML Handbook, Oxford University Press, 1990.

[HARV] Bowman, C. Mic et.al. "Harvest: A Scalable, Customizable Discovert and Access System" Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado - Boulder, http://harvest.cs.colorado.edu/

[HORN] Horny, Karen L., "Minimal-level cataloging: A look at the issues- A symposium", Journal of Academic librarianship, Vol. 11, pp. 332-334.

[INFO] InfoSeek Home Page, http://www.infoseek.com/

[KAHL] Kahle, Brewster, "An Information System for Corporate Users: Wide Area Information Servers", Thinking Machines Technical Report TMC-199, April 1991. Also in On-line Magazine, August 1991 and ftp://ftp.wais.com/pub/wais-inc-doc/txt/WAIS-Corp.txt

[KOST] Koster, M. "ALIWEB(Archie Like Indexing the WEB)", http://web.nexor.co.uk/aliweb/doc/aliweb.html

[LUCA] Peter, Lucas. "Refinement of the HEPAR Expert System: Tools and Techniques", In Artificial Intelligence in Medicine, 6(2); pp. 175–188, 1994.

[MARC] Library of Congress, "MARC manuals used by the Library of Congress", American Library Association, Chicago, 1969.

[MAUL] Mauldin, Michael L., "Measuring the Web with Lycos", Poster Proceeding of the third International WWW Conf., Darmstadt, April 1995, pp. 26-29. also http://lycos.cs.cmu.edu/

[MCBR] McBryan, Oliver A., "World Wide Web Worm", http://www.cs.colorado.edu/home/mcbryan/WWWW.html

[NIKO] Nikos, World Wide Web Index, Rockwell Network Systems http://www.rns.com/

[PETE] Petersen, Toni, Molholt, Pat (ed), "Beyond the book: extending MARC for subject access", G.K. Hall, Boston, MA, 1990.

[RBSE] Eichmann, David, "RBSE Program", http://rbse.jsc.nasa.gov/eichmann/rbse.html, http://rbse.jsc.nasa.gov/eichmann/urlsearch.html

[RFC1738] "Uniform Resource Locators(URL)", T. Berners-Lee, L. Masinter, M. McCahill:, can be obtained via anonymous FTP from anyone of: ds.internic.net, nis.nsf.net, src.doc.ic.ac.uk, munnari.oz.au and a number of other sites.

[ROSS] Ross, Rayburn M., West, Linda, "MLC: A contrary viewpoint", Journal of Academic Librarianship, Vol. 11, pp.334-336

[RHEE] Rhee, Sue, "Minimal-level cataloging: Is it the best local solution to a national problem? ", Journal of Academic librarianship, Vol. 11, pp.336-337, 1986.

[SHIN] Shinghal, R. "Formal Concepts in Artificial Intelligence", Chapman & Hall, London, U.K., co-published in U.S. with Van Nostrand, New York, 1992.

[THAU] Thau, R., "SiteIndex Transducer", http://www.ai.mit.edu/tools/ site-index.html

[WEBC] WebCrawler, http://www.biotech.washington.edu/WebCrawler/WebQuery.html

[WWWC] World Wide Web Catalog, http://cui_www.unige.ch/

[YAHO] http://www.yahoo.com/search.html

The Document Type Definition

The Document Type Definition (DTD)uses a set of rules in the description of the document content. Some of these are listed below:

- An element followed by the symbol "?" is optional and could occur zero or one time.
- An element followed by the symbol "*" can occur zero or more times. This symbol is called the *optional and repeatable occurrence indicator* [GOLD].
- An element followed by the symbol "+" is required and can occur one or more times. This symbol is called the *required and repeatable occurrence indicator* [GOLD].
- The percent sign is used to define a subset of elements before its first reference. Example:

  "%ROLE" is the reference to the entity named ROLE. Each entity should be declared between tags as follows:

```
<!ENTITY % ROLE  "Author | Artist | Co-author | Corporate Entity |
                  Editor |Designer |Programmer | Publisher | Other">
```

```
<!-- Parameterizable list -->
<!ENTITY % ROLE  "Author | Co-author | Editor | Artist | Corporate Entity | Designer |
                  Programmer | Publisher | Other">
<!ENTITY % DOM-ID  "FTP | ISBN | ISSN | Gopher | HTTP | SHN | UAS | URN | Other">
<!ENTITY % DOM-CLASS "Legal | Security | Nature | Other">
<!ENTITY % DOM-CVRG  "Audience | Geographical Coverage | Spatial Coverage | Epoch | Other">
<!ENTITY % DOM-SYSRQ  "Hardware | Network | Software | Other">
<!-- Element list -->
<!ELEMENT Semhdr --
         (Title, Alt-title?, Subject+, Language?, Char-Set?, Author+, Keyword+, Dates,
          Version?, Supersede?, Classification*, Coverage*, Sysreq*, Genre?, Source*,
          Cost?, Abstract?, Annotation*, Control>
<!ELEMENT Title           --          (#PCDATA) >
<!ELEMENT Alt-title       --          (#PCDATA) >
<!ELEMENT Subject         -0          (EMPTY)   >
<!ELEMENT Language        --          (#PCDATA) >
<!ELEMENT Char-Set        --          (#PCDATA) >
<!ELEMENT Author          -0          (EMPTY)   >
<!ELEMENT Keyword         --          (EMPTY)   >
<!ELEMENT Identifier      --          (#PCDATA) >
<!ELEMENT Dates           --          (Created, Expiry?, Updated?) >
<!ELEMENT Created         --          (#PCDATA) >
<!ELEMENT Expiry          --          (#PCDATA) >
<!ELEMENT Updated         --          (#PCDATA) >
<!ELEMENT Version         --          (#PCDATA) >
<!ELEMENT Supersede       --          (#PCDATA) >
<!ELEMENT Classification  --          (#PCDATA) >
<!ELEMENT Coverage        --          (#PCDATA) >
<!ELEMENT Sysreq          --          (#PCDATA) >
<!ELEMENT Genre           --          (#PCDATA) >
<!ELEMENT Source          --          (#PCDATA) >
<!ELEMENT Abstract        --          (#PCDATA) >
<!ELEMENT Annotation      --          (#PCDATA) >
<!ELEMENT Cost            --          (#PCDATA) >
<!ELEMENT Control         --          (#PCDATA) >
```

Figure 3: DTD for Semantic Header

```
<!-- Attribute List -->
<!ATTLIST Subject        "General"            CDATA            #REQUIRED
                         "Sublevel1"          CDATA            #IMPLIED
                         "Sublevel2"          CDATA            #IMPLIED  >
<!ATTLIST Author         "Role"               (%ROLE;)         #REQUIRED
                         "Name"               CDATA            #REQUIRED
                         "Organization"       CDATA            #IMPLIED
                         "Address"            CDATA            #IMPLIED
                         "Phone"              CDATA            #IMPLIED
                         "Fax"                CDATA            #IMPLIED
                         "EMail"              CDATA            #IMPLIED  >
<!ATTLIST Identifier     "(%DOM-ID)"          CDATA            #REQUIRED >
<!ATTLIST Classification "(%DOM-CLASS;)"      CDATA            #REQUIRED >
<!ATTLIST Coverage       "(%DOM-CVRG;)"       CDATA            #REQUIRED >
<!ATTLIST Sysreq         "(%DOM-SYSRQ;)"      CDATA            #REQUIRED >
<!ATTLIST Annotation     "Annotation"         CDATA            #IMPLIED
                         "Signature"          #PCDATA          #IMPLIED >
<!ATTLIST Genre          "Form"               CDATA            #IMPLIED
                         "Size"               CDATA            #IMPLIED >
<!ATTLIST Cost           "Currency"           CDATA            #IMPLIED
                         "Amount"             #PCDATA          #IMPLIED >
<!ATTLIST Control        "Account"            CDATA            #IMPLIED
                         "Password"           #PCDATA          #IMPLIED >
```

Figure 4: DTD for Attribute List for Semantic Header of Figure 3

- Data types and keywords:

  #PCDATA : zero or more parsed data characters

  CDATA : character data

  EMPTY : empty content

The DTD for the Semantic Header is given in Figure 1 below.

The atribute lists for the elements Subject, Author, Identifier, Classification, Coverage, Sysreq, Annotation, Genre, Cost and Control of Figure 3 are given in Figure 4.