Indexing and Searching Virtual Libraries[1]

(c) Bipin C. **Desai**
Department of Computer Science
Concordia University
7141 Sherbrooke St. W
Montreal, H4B 1R6
bcdesai@cs.concordia.ca
http://www.cs.concordia.ca/~faculty/bcdesai/

Keywords: Bibliographic record, Indexing, Searching, Discovery, Dublin Metadata Elements List, Core Elements, Semantic header, Database system, Expert system, Searching

Abstract

It is well known that selectivity leaves a lot to be desired in searching for information resources on the Internet with existing search systems[DESA4]. This has prompted a number of researchers to turn their attention to the development and implementation of models for indexing and searching information resources on the Internet. In this white paper[2] we examine briefly the results of a simple query on a number of existing search systems and then discuss two proposed index metadata structures for indexing and supporting search and discovery: the Dublin Core Elements List and the Semantic Header.

Introduction

Access to relevant information is one of the most important requirements of all human endeavours. This need has been recognized and has resulted in the continuing effort to describe and organize information so as to facilitate its expected discovery and ready 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 the Web[BERN, BERN3] in particular. This is now becoming the accepted method of disseminating and sharing information resources in hypermedia. At this time a number 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 speciality newsletters are available.

A number of index generation systems and related search systems are currently available on the Internet[DEBR, EMTA, FLET, KAHL, KOST, MAUL, MCBR, SEAR, THAU, WEBC, WWWW, YAHO]. Some of these are manually generated indices(Aliweb[KOST], CUI W3 Catalog[WWWC], GNA Meta-Library[GNAM], DA-CLOD[DACL]) while others

---

[1] The Semantic Header is component of the CINDI subsystem - a part of the CUILT Project for Developing a Virtual Library Prototype.
This work is supported in part by a grant from Seagram's Fund for Academic Innovation.
[2] CIC Forum: America in the Age of Information, Bethesda, MD, July 1995 .

are generated by robots (Harvest[HARV], Lycos[MAUL], Nikos[NIKO], Yahoo[YAHO], Web Crawler[WEBC]). 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 claims to have indexed nearly 4 million documents has only partial success in locating all relevant documents[DESA4][3].

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	18
GNA Meta Lib.	none	-	-	24
Harvest	none	-	-	24
Lycos	231	2	222	17
Nikos	none	-	-	24
W3 Catalog	none	-	-	24
WebCrawler	7	3	0	20
Yahoo	none	-	-	24

Table 1 Search statistics for using the search term Bipin (AND) Desai

The Problem

The purpose of indices and bibliographies (called secondary information) is to inventory the primary information and 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 and [DESA4] illustrates that there is a need for the development of a system which allows better controlled 'search for' and 'access to' resources available on the

[3] The list of URLs of document known to contain the search terms is given in [DESA4]. The tests were done on June 3, 1995. Result may not be identical if the tests are repeated due to the possible discovery of the missing documents by the index systems involved. **All** documents in [DESA4] existed *well* before the test date.

Internet. With the current plethora of index services and search systems, most users are lost. However, 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 structure and user interface.

This phenomena has been observed in distributed information systems which, even though under control of a single administrative unit, create multiple problems typically caused by differences in semantics and representation, and incomplete and incorrect data dictionaries (cataloging) [DESA]. These problems would be magnified manyfold in any distributed information system which tries to integrate the resources offered by indexing and search systems over the Internet. It is important, also, to avoid problems encountered in a library system where, in spite of the fact that while the same cataloging system[4] is used, the same item may be differently catalogued/classified in two different libraries.

Such problems could be avoided by starting with a standard index structure and building a bibliographic system using standardized control definitions. Such definitions could be built into the knowledgebase of expert system-based index entry and search interfaces. Furthermore, there must be a mechanism to revise index information as the resource changes over time. Finally, annotation of a resource by independent users should be allowed.

With the increasing amount of information on the Internet, it is difficult to follow such a traditional centralized approach due to the enormous number of resources involved and hence the time and cost. Consequently, the indexing system should be distributed and accessible to providers as well as users of the Internet. In a distributed system such as the Internet, it is natural to have the providers of resources, prepare and enter the bibliographic information about each resource using a standardized index scheme. The entry system should be a distributed system and the index should be recorded in a distributed database. Finally, a search system to facilitate locating and retrieving appropriate information from this database is required.

Whereas the index entry and search systems (clients) could be located locally at the providers and users of information resources respectively, the bibliographic database systems(servers) should be distributed and replicated at a number of regional nodes for enhanced availability and response. The entry and search systems have to be supported by an easy-to-use graphical interface for entering the index information and accessing it. These systems should incorporate the expertise and knowledge of expert cataloguers and reference librarians with a help system to guide the user at all steps. The search system should, in addition, provide appropriate feedback indicating the number of hits for each search, and support access to the relevant resources. The navigation of database and resource nodes and the protocols and filters used would be selected by the system, thus facilitating the task of the user. The purpose is to provide uniform access to all resources, as is done in a centralized information system through the intermediary of an expert system analyst.

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 informa-

[4] Libraries use a number of basic catalogue systems such as Library of Congress, Dewey Decimal and MARC. Even among MARC there are slight differences as in LCMARC and CANMARC.

tion. However, such systems are designed for professional catalogers and many of the elements included in them, though useful, are beyond the level of familiarity of most providers or users of information.

In the following sections, we present two recent initiatives for allowing suppliers of resources to prepare well thought out catalog information for their resources.

Dublin Metadata Workshop

The Metadata Workshop was held from March 1 to March 3rd, 1995 in Dublin, OH[DESA3]. It was open "by invitation only" to a number of people actively involved in one or another aspect of the Digital or Virtual Library project, primarily in North America. The intent of the meeting was to try to work towards the definition of a minimum common set of elements for Network Object. The workshop brought together select professionals from computer science, library science, professional librarians, as well as professionals involved in: on-line information services; abstracting, cataloging and indexing; imaging and geospatial data; and museums and archives. The main objective was to address the problem of cataloging network resources with adoption, extension or modification of current standards and protocols to facilitate their discovery and access. A list of participants and their affiliations is given in [DESA3].

The goals of the workshop were: to achieve a consensus on a set of core data elements for document-like objects(DLO); to map these and related elements to accepted standards; and to devise an extension scheme for registering other types of network objects. It was hoped that the workshop, which was preceded by a continuing discussion via a restricted discussion list-server, would promote common understanding of the needs of the various communities being served by the network. The approaches and solutions proposed by these communities, and their strengths and weakness would be recognized in developing a minimum core element set.

In spite of the objective stated in the workshop invitation, many participants had no expectation of coming up with a comprehensive list of data elements. However, there was hope that some categorization and definition of the concept necessary for supplier and user of information would emerge. The other thesis that was repeated in the on-line discussions that preceded the workshop was that in spite of divergence in metadata elements from one catalog to another, there should be a mechanism to provide inter-operability. It was recognized that a metadata element list that would work for everything would be difficult to achieve.

Many participants did recognize the futility of an exhaustive standard and rather wanted to determine a non- exhaustive list of characteristics of network resources and users, as well as the method of their use. It was recognized that the value of information is enhanced if it is represented in an 'application neutral' manner. The same can be said about representing the semantic content of information about information.

Another concern that was expressed was that the core elements should serve as a nucleus for future enhancements. This would be a difficult problem were we to have no idea about future requirements. Another concern raised was the intended use of the metadata. Different usage may impose different constraints and need different sets of core elements. Consequently, the following assumptions were made to develop a consensus to arrive at a minimum set of core data elements.

Assumptions

1. The elements are intended to describe a **Document Like Object** (DLO: Loosely defined as an object which is similar to an object considered to be a document); no attempt is made to assert the suitability of this element set for all possible object types.
2. Common (or core) element set: These are metadata elements that apply to most/many DLOs.
3. The elements of the core are chosen to support resource discovery: recognizing from these elements enough about the target objects to know if they do or do not meet the current information requirements.
4. All elements of the metadata are repeatable.
5. All elements are optional.
6. All elements describe the DLO itself with the exception of the SOURCE element, which can be thought of as a recursive instance of the entire record, except that it applies to an object from which the electronic record is derived.
7. The Core elements are intended to describe intrinsic characteristics of the DLO... thus, transactional data, archival status, and copyright characteristics (as well as others) are not included in this set.
8. No assumption is made concerning whether the DLOs are network accessible or specifically electronic.
9. The core element set assumes an arbitrarily complex hierarchy.
10. Elements not included in the core set are not specifically excluded.

Note: Any implementation will require an extensibility mechanism to include other elements, either of local significance or pointers to other established element sets (MARC, TEI, etc.)

THE PROPOSED ELEMENT LIST:

The following elements emerged as the ones required in a minimum set. They were dubbed the **`Dublin Metadata Element List(DMEL)`**.

subject: words or phrases indicative of the information content. If the value comes from a controlled vocabulary, the SCHEME sub-element is used to indicate the particular vocabulary.

title: the title, name or short description of the object.

author: the name of the creator of the content: order of names according to the culture.

otherAgent: the name of any other entity responsible for the content of the object: the role sub-element describes the responsibility.

publisher: the name of the entity responsible for making the object available.

date: the date of publication.

identifier: a character string or a number used to distinguish this object from other objects: a SCHEME sub-element identifies the authority.

object-type: conceptual description of object.

form: physical, logical, or encoding characteristics.

relation: important known relationship to other objects; the TYPE sub-element describes the nature of the relationship; the SCHEME sub-element identifies the notation used to identify the related object(s).

language: natural language of the content of the object; the SCHEME element identifies the controlled vocabulary edition: a free-text identification of the version.

source: object from which derived; contains a nested object description.

coverage: characterizes parameters to specify the audience, or the time or space (instance or span).

Comments on DMEL

Designating all elements of the DMEL as optional may create a problem; a minimum set should be required.

The user may need some extrinsic characteristics of the resource such as its cost and hence should be included (though optional). All documents of any permanence should be archived and accessible.

Subject element should be made up of two sub-fields, a schema and an hierarchical subject field which includes sub-subject and sub-sub-subject. The classification scheme used (authority) could be specified by the schema field entry specifying the cataloging schema. The hierarchical subject field entries must be from the same schema.

For non-titled objects, we need an algorithm to insert a(n alternate) title. For example the non-titled resources such as satellite data can have a cooked up title using satellite name, time, date, position, camera orientation, and filter or frequencies etc.

Relation and source have similar semantic implications and could be described using a relationship subfield with the identifier of the object to which it is related. An optional sub-field may be used to provide annotation or useful information.

In addition to the language of the DLO, the (natural) language used to specify the elements of the metadata and the character set used have to be specified. Furthermore, elements such as an abstract and annotation are **missing**. The former, generated by a human or other agent gives a capsule idea of the contents of the DLO. The latter is a placeholder for additional details regarding the DLO by responsible agents or other users.

Semantic Header

Professional cataloguers have found the need for elements similar to those in DMEL in most indexing applications. This dictates that they must be included in most indexes for information resources. The dependence on titles as the most commonly used 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 add annotations, keywords or key phrases to indicate the actual content.

Accuracy or quality of a document can be indicated by including reviewers' opinions. However, such opinions 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 selected from scanning the text. Reference librarians and library users tend to use such annotated bibliographies to help choose among competing sources. Fortunately, for an on-line index system as proposed in CINDI[DESA2], 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[DESA1] 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(April 1994). Since then, it has been extended to other resources accessible directly on the Internet.

The structure of the index is similar to the ones used for most library indices and include other information deemed useful for on-line systems. The semantic header may be considered as an application of SGML[GOLD]. 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. Figures 1 through 3 below give the DTD for the Semantic Header.

The intent of the semantic header is to include those elements that are most often used in the search for an information resource. Since the majority of searches begin with a title, name of one of the authors (70%), subject and sub-subject (50%)[Katz], we have made the entry of these elements mandatory in the semantic header. The abstract and annotations are, as well, relevant in deciding whether or not the resource would be useful; these items are also included. The elements of the semantic header are described briefly below:

```
<!-- Parameterizable list -->
<!ENTITY % SE_SBJCT  '(General,(SubLevel1, SubLevel2?)?)+'>
<!ENTITY % SE_RA '(Role, Name, Organization?, Address?, Phone?, Fax?, EMail?)+'>
<!ENTITY % ROLE  "Author | CoAuthor | Editor | Artist | Composer | Corporate Entity |
                  Designer | Programmer | Publisher | Sponsor | Translator | Other">
<!ENTITY % SE_KW   'Kw+'>
<!ENTITY % SE_ID '(IdDomain, IdValue)+'>
<!ENTITY % DOM_ID  "FTP | ISBN | ISSN | Gopher | HTTP | LCCN | SHN | UAS | URN | Other">
<!ENTITY % SE_DT '(DSchema, Created, Expiry?, Updated?)'>
<!ENTITY % DATE_SCHEMA "YYYY | YYYY-MM-DD | Other" >
<!ENTITY % SE_VR '(Current,  Supersede?)?'>
<!ENTITY % SE_CLASS '(ClassDomain, ClassValue)*'>
<!ENTITY % DOM_CLASS "Legal | Security | Nature | Other">
<!ENTITY % SE_CVRG '(CovDomain, CovValue)*'>
<!ENTITY % DOM_CVRG  "Audience | Geographical Coverage | Spatial Coverage |
                      Epoch | Other">
<!ENTITY % SE_SYSRQ '(SysDomain, SysValue+)*'>
<!ENTITY % DOM_SYSRQ  "Hardware | Network | Software | Other">
<!ENTITY % SE_GNR '(Form, Size)*'>
<!ENTITY % SE_SRC '(Relationship, IdDomain, IdValue)*'>
<!ENTITY % RELATIONS  "Contains| ContainedIn | ContinuedFrom | ContinuedTo |
                          DerivedFrom | IndexOf | IndexedIn |
                           PartOf | PrecededBy | FollowedBy | Other">
<!ENTITY % SE_COST '(Currency, Amount)*'>
<!ENTITY % SE_ANN '(Annotation, Signature)*'>
<!ENTITY % SE_CNTRL '(Account, Password)'>
```

Figure 1 DTD for Semantic Header: Entitiess

Title, Alt-title

The first field of the semantic header is the title[5] of the resource. It is a name given to the resource by its creator(s) and is a required field. In the formal definition it is enclosed within the tags beginning with <title> and terminated by </title>. The title could include the sub-title, as is done in many cataloging systems. The alternate title field is enclosed by the tags <alt-title>, </alt-title> and used to indicate an "official" secondary title or an alternate title of the resource. Whereas the element title is a required element, the alternate title is optional.

```
<!--                              Element list                                  -->
<--       Element     Minimization   Value           Default   -->

<!ELEMENT SemHdr - -  (Title, AltTitle?, Subject, Language?, CharSet?, RespAgent,
                      Keywords, Identifier, Dates, Version, Classification, Coverage,
Sysreq,
                      Genre, Source, Cost, Abstract?, Annotation, Control)       >
<!ELEMENT Title           - -        CDATA          #REQUIRED   >
<!ELEMENT AltTitle        - -        CDATA          #IMPLIED    >
<!ELEMENT Subject         - -        (% SE_SBJCT;)              >
<!ELEMENT Language        - -        CDATA          #IMPLIED    >
<!ELEMENT CharSet         - -        CDATA          #IMPLIED    >
<!ELEMENT RespAgent       - -        (% SE_RA;)                 >
<!ELEMENT Keywords        - -        (% SE_KW;)                 >
<!ELEMENT Identifier      - -        (% SE_ID;)                 >
<!ELEMENT Dates           - -        (% SE_DT;)                 >
<!ELEMENT Version         - -        (% SE_VR;)                 >
<!ELEMENT Classification  - -        (% SE_CLASS;)              >
<!ELEMENT Coverage        - -        (% SE_CVRG;)               >
<!ELEMENT Sysreq          - -        (% SE_SYSRQ;)              >
<!ELEMENT Genre           - -        (% SE_GNR;)                >
<!ELEMENT Source          - -        (% SE_SRC;)                >
<!ELEMENT Cost            - -        (% SE_COST;)               >
<!ELEMENT Abstract        - -        CDATA          #IMPLIED    >
<!ELEMENT Annotation      - -        (% SE_ANN;)                >
<!ELEMENT Control         - -        (% SE_CNTRL;)              >
```

Figure 2 DTD for Semantic Header: Elements

Subject

The subject and sub-subjects of the resource are indicated in the next field which is a repeating group (a multi-part field with one or more occurrences of items in the group). All resources must have at least one occurrence for this field.

Language, Character set

The character set used and the language of the resource is given in the next two optional fields.

Author and other responsible agents

The details about the author(s) and/or other agent(s) responsible for the resource is given in the next repeating group[6]. The sub-fields are: role[7] of the agent, name,

---

[5] Title for non document like resources may require some creativity. For instance the title of a satellite image could be generated from the name of the satellite, its location, date, time, cameras, frequencies, filters, etc.

[6] For resources such as satellite image, the agent may be the agency controlling the satellite or the satellite itself.

[7] Typical values for role of the agent could be author, co-author, designer, editor, programmer, creator, artist, corporate entity, publisher, etc.

organization, address, phone and fax numbers, and e-mail address. All sub-fields save the name are optional, except in the instances of corporate entities in which case the organization must be given. By using the role sub-field and giving it appropriate value, semantics for agents such as editor or publisher are incorporated in this repeating group.

```
<--        Element      Minimization   Value              Default -->

<!ELEMENT  General          - O      CDATA              #REQUIRED >
<!ELEMENT  Sublevel1        - O      CDATA              #IMPLIED  >
<!ELEMENT  Sublevel2        - O      CDATA              #IMPLIED  >
<!ELEMENT  Role             - O      (%ROLE;)           #REQUIRED >
<!ELEMENT  Name             - O      CDATA              #REQUIRED >
<!ELEMENT  Organization     - O      CDATA              #IMPLIED  >
<!ELEMENT  Address          - O      CDATA              #IMPLIED  >
<!ELEMENT  Phone            - O      CDATA              #IMPLIED  >
<!ELEMENT  Fax              - O      CDATA              #IMPLIED  >
<!ELEMENT  EMail            - O      CDATA              #IMPLIED  >
<!ELEMENT  Kw               - O      CDATA              #IMPLIED  >
<!ELEMENT  IdDomain         - O      (%DOM_ID;)         #REQUIRED >
<!ELEMENT  IdValue          - O      #PCDATA            #REQUIRED >
<!ELEMENT  DSchema          - O      (%DATE_SCHEMA;)    #REQUIRED >
<!ELEMENT  Created          - O      #PCDATA            #IMPLIED  >
<!ELEMENT  Expiry           - O      #PCDATA            #IMPLIED  >
<!ELEMENT  Updated          - O      #PCDATA            #IMPLIED  >
<!ELEMENT  Current          - O      CDATA              #IMPLIED  >
<!ELEMENT  Supersede        - O      CDATA              #IMPLIED  >
<!ELEMENT  ClassDomain      - O      (%DOM_CLASS;)      #REQUIRED >
<!ELEMENT  ClassValue       - O      #PCDATA            #REQUIRED >
<!ELEMENT  CovDomain        - O      (%DOM_CVRG;)       #REQUIRED >
<!ELEMENT  CovValue         - O      #PCDATA            #REQUIRED >
<!ELEMENT  SysDomain        - O      (%DOM_SYSRQ;)      #REQUIRED >
<!ELEMENT  SysValue         - O      CDATA              #REQUIRED >
<!ELEMENT  Form             - O      CDATA              #IMPLIED  >
<!ELEMENT  Size             - O      CDATA              #IMPLIED  >
<!ELEMENT  Relationship     - O      (%RELATIONS;)      #REQUIRED >
<!ELEMENT  IdDomain         - O      (%DOM_ID;)         #REQUIRED >
<!ELEMENT  IdValue          - O      #PCDATA            #REQUIRED >
<!ELEMENT  Currency         - O      CDATA              #IMPLIED  >
<!ELEMENT  Amount           - O      #PCDATA            #IMPLIED  >
<!ELEMENT  Annotation       - O      CDATA              #IMPLIED  >
<!ELEMENT  Signature        - O      CDATA              #IMPLIED  >
<!ELEMENT  Account          - O      CDATA              #IMPLIED  >
<!ELEMENT  Password         - O      SECRET             #IMPLIED  >
```

Figure 3 DTD for Semantic Header: Sub-elements

Keyword

The list of keywords is included in this field.

Identifier

The next element is a repeating group for recording the identifiers of the resource. Each occurrence of this group consists of two sub-fields: one for the domain and the other for the corresponding value.

The domain could be an accepted or standardized coding scheme issued by an appropriate authority such as ISBN, ISSN, URL(FTP, GOPHER, HTTP)[BERN1], or URN[RFC1737] etc., and the value contains the corresponding coded identifier. Since a resource in electronic form may be accessible from one or more sites there could be one or more entries for the same domain such as URL. The URN field gives the unique name of the resource, if any. This name may be used instead of a location (URL) if the item is

likely to move or is accessible from multiple locations[8]. The identifier(s) can be used to locate the resource.

In the absence of an accepted standard for URN, we use an alternate name, called Semantic Header Name(SHN). The SHN is derived by concatenating the following required elements in the semantic header: the title, name of first author (or name of organization, if the resource is attributable to a corporate or organizational entity), first subject, and creation date. The string generated is prefixed by the initial location of the resource and suffixed with and an optional system-generated integer number for possible disambiguation. With this scheme, the user supplied elements in the SHN, with a very small probability, may map to more than one resource. If multiple hits are encountered during a search based on user supplied elements of the SHN, the system would inform the user of the "collision". The user could then select the appropriate resource index entry by perusing the other elements recorded in the semantic header.

The identifier entry in the semantic header may also contain an entry for an archive site. The domain value UAS (universal archive site) is used to indicate the archive site for the resource. It is expected that the resource will exist at this site beyond its expiry date, if any. Of course, the site itself is guaranteed to exist beyond the life of any resource. It is envisaged that the archive site could be an independent resource provider. Examples of such traditional resource providers that would be feasible archive sites for the resource are the national libraries such as the Library of Congress in U.S., British Library, National Library and CISTI in Canada. However, private, for profit, corporations could be alternate sites for archiving resources. Archiving would provide an anchor for the otherwise ephemeral nature of some resources on the network. Since the archive site may not be known when the semantic header is first registered, the system would support update operations in which existing entries could be modified. Other update operations such as modification of addresses, URLs etc., would also be supported.

Dates

The dates of creation(required), expiry and update, if any, are given next. Any updates made are indicated by a system generated date.

Version

The version number, and the version number being superseded if any, are given in these optional elements.

Classification

The intended classification is indicated in the next optional repeating group. It consists of a domain (nature of resource, security or distribution restriction, copyright status, etc.) and the corresponding value.

Coverage

The coverage is indicated in the next optional repeating group. It consists of a domain (target audience, coverage in a spatial and/or temporal term, etc.) and the corresponding value.

---

[8] The idea of the semantic header is to provide bibliographic information about resources and by including both the SHN/URN and a list of URLs it also provides a mapping from SHN/URN to URLs.

System Requirements

A list of system requirements such as hardware and software required to access, use, display or operate the resource is included in the semantic header as an optional repeating group. It consists of a domain of the system requirements (possible values are: hardware, software, network, protocols, etc.,) and the corresponding exigance.

Genre

This optional element is used to describe the physical or electronic format of the resource. It consists of a domain (type of representation or form which in the case of a file could be its format such as ASCII, Postscript, TeX, GIF, etc.,) and the corresponding value or size of the resource.

Source/Reference

The relationship of the resource to other resources may be indicated by the optional repeating group. It contains the relationships, domains and identifiers of related resources. A related object may be used in deriving the resource being described, or it may be its sub/super components. Such information, is usually found in the body of a document-like resource. However, this optional group permits an option for this type of resource and an opportunity to register it for resources of other formats.

Cost

In the case of a resource accessible for a fee, the cost of accessing it[9] is given next. It consists of a currency and the cost for accessing the resource.

Abstract and Annotations

The abstract and annotations are given in the next fields. The abstract is provided by the author of the resource; the annotations are made by the author and/or independent users of the resource and include their identities along with their digital signatures. Once registered, the annotations *cannot* be modified.

Control

The last set of items in the semantic header is that of the control items such as the account to which credits are to be made for charges for accessing the resource, encoded passwords or the digital signature of the provider of the resource. Any change to the update-able part of the semantic header requires the password or digital signature. Another control piece of information is the digital signature of the resource itself. This may be used to authenticate the resource when it is retrieved through a semantic header. It is assumed that there is a mechanism to access the resource's digital signature.

Importance of Metadata for Indexing and Searching and Discovery

Metadata is the information which records the characterization and relationship of the source data. It helps to provide succinct information about the source data which may not be recorded in the source itself due to its nature or an oversight.

[9] Such cost could change over time and require updating.

In this white paper, we limit our discussion to the importance of metadata for indexing to support subsequent search and discovery operations by future users. Presently, users are able to search for and obtain, the required information, after a number of trials with various indexing services. However, unless the challenges outlined below are met, this may not be the case in the year 2010.

Due to the sheer volume of data in the emerging information infrastructure, search and discovery would become difficult without some well thought out discovery mechanism built around adequate metadata. Consider what would happen if one had to search for a specific volume from the LC if its entire collection were piled together, helter-skelter, in a darkened hanger. The task becomes even more daunting if we were not looking for a specific volume but for a volume which dealt with such-and-such topic. The problem with current automatically generated index databases is their inadequate semantic information. Yet, it is evident that professional cataloging of the ever-growing information resources, would be prohibitively expensive. Thus, the design of adequate metadata to describe and establish the semantic contents of resources and to establish their semantic dependencies on other resources is of utmost importance. This, along with a registering system, would establish a basis for later search and discovery.

Metadata would provide an instrument to describe the semantic content of a resource. Such metadata is better suited to supporting discovery than the resource itself. In many cases the resources themselves may not be able to provide the semantic dependencies or it would be computationally too expensive to do so. (For example how does one conclude that a given program code is used to provide computation of consumer loan payments without analyzing the program.) Metadata, for instance, facilitates the cataloging of resources such as audio, computer programs, services, images and videos. This becomes important when the resource itself is not as easily accessible as the index

Another reason for using metadata and extracting salient features of a resource is to support retrieval by content. Automatic processing of the contents of a source by extractors have been done on an ad-hoc basis but have been found to be unreliable. A case in point is the promise of NLP not quite realized. Approximations such as WAIS have been useful but have also shown that relevancy measures derived using frequencies, proximity etc. may not always be meaningful.

Metadata could also be used to express semantic dependencies which are inherent in a collection of objects. This means that the structure of the objects could be expressed using metadata as their surrogates and the actual sources could be separated from their metadata. This simplifies the storage of the resources and allows for the recognition of redundancies. Extracting such semantic dependencies in metadata allows for search based on the contents of multimedia resources.

Initial query processing could be done on the metadata and thus avoid access to most of the resources and the possibility of their computationally bound interpretation. This becomes more advantageous when there are costs (time, money, network bandwidth and overloading) involved in accessing resources. The cost of accessing metadata would be much smaller than the cost of accessing the resource. Query processing would be supported by statistics, and an expert system to help formulate queries as is done by a research librarian.

Appropriately constructed metadata could support query based on contents as well as traditional query based on items such as title, author, subject, etc. This means that the structure of the objects could be expressed using metadata as their surrogates and the

actual sources could be separated from their metadata. This simplifies the storage of the resources.

The challenge of the coming information age in the area of metadata can be summarized as: defining an extensible metadata structure; automatic and semi-automatic (human assisted) extracting of metadata from resources; designing of a distributed indexing system; designing of a query language to support discovery and provide location transparency; designing of expert based resource registering and searching systems; and designing of an intuitive graphical user interface to interact in the discovery process.

Conclusions

Current index systems are based on harvesting the network for new documents. Such documents are retrieved and their contents used to provide terms for the index. The big disadvantage with this scheme is the unreliability of the index entries produced and the lack of an authentic abstract for the item. The current Dublin Metadata Element list also suffers from the absence of the abstract. Current index schemes are relevant for resources of limited protocols and are not applicable to other resources. Another problem with some of the robot-based approaches is the unnecessary traffic on the network and lack of cooperation and sharing among different systems. Finally, the infeasibility of the existing approach becomes clear as more and more providers of information would require payments.

In the system based on the Semantic Header, the provider of the resource is the one who prepares the index information. Consequently, such index entry would be more reliable than one derived by a third party or by simply scanning a document. The inclusion of an abstract in the index entry enables the provider of the resource to highlight the nature of the subject. With a fee involved, users would not be inclined to retrieve resources with irrelevant titles. The Semantic Header provides additional details about the resource and allows users to make better informed decisions regarding the relevance of the source resource.

The system provides an expert system-driven graphical interface for the provider of the resource to produce an index entry, and have this entry entered in the index database. The expert system provides help in choosing appropriate terms for index entries such as subject, sub-subject, keywords etc. It also is responsible for verifying the consistency of the index entry and the accessibility of the resource and then for posting the index entry to the index database.

Lastly, the index database contains a number of control entries for the resource. Control entries are items such as the size of the resource, the password for authenticating subsequent updates of the index entry, and a list of annotations made about the resource by independent users.

References

[BERN] Berners-Lee, T., Cailliau, R., "WorldWideWeb: Proposal for a HyperText Project"
http://info.cern.ch/hypertext/WWW/Proposal.html
[BERN1] Berners-Lee, T. "UR* and The Names and Addresses of WWW objects",
http://info.cern.ch/hypertext/WWW/Addressing/Addressing.html
see also RFC 1738,

[BERN2] Berners-Lee, Tim, Connolly, "Hypertext Markup Language, Internet working draft", http://info.cern.ch/hypertext/WWW/MarkUp/HTML.html
[BERN3] 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.
[DACL] http://schiller.wustl.edu/DACLOD/daclod
[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.
[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., Pollock, Richard, "MDAS: A Heterogeneous Distributed Database Management System", Information and Software Technology, January 1992, Vol. 34-1, pp. 28-41.
[DESA1] Desai, Bipin C., "Cover page aka Semantic Header", July 1994, http://www.cs.concordia.ca/semantic-header.html, revised version, August 1994, http://www.cs.concordia.ca/~faculty/bcdesai/semantic-header.html
[DESA2] Desai, Bipin C., "The Semantic Header and Indexing and Searching on the Internet", February 1995, http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html
[DESA3] Desai, Bipin C., "Report of the Metadata Workshop, Dublin, OH", March 1995, http://www.cs.concordia.ca/~faculty/bcdesai/metadata-workshop-report.html .
[DESA4] Desai, Bipin C., "Internet Indexing Systems vs List of Known URLs", June 1995 URL:http://www.cs.concordia.ca/~faculty/bcdesai/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.
[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] 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.
[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
[KATZ] William A. Katz, "Introduction to Reference Work", Vol. 1-2 McGraw-Hill, New York, 1987
[KOST] Koster, M. "ALIWEB(Archie Like Indexing the WEB)", http://web.nexor.co.uk/aliweb/doc/aliweb.html
[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 see http://lycos.cs.cmu.edu/
[MCBR] McBryan, Oliver A., "World Wide Web Worm", http://www.cs.colorado.edu/home/mcbryan/WWWW.html
[NIKO] http://www.rns.com/
[PETE] Petersen, Toni, Molholt, Pat (ed), "Beyond the book: extending MARC for subject access", G.K. Hall, Boston, MA, 1990.
[RFC1357] "A Format for E-mailing Bibliographic Records", D. Cohen.: 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.
[RFC1737] "Functional Requirements for Uniform Resource Name", K. Sollins, L. Masinter: pl. see RFC1357 above.
[RFC1738] "Uniform Resource Locators(URL)", T. Berners-Lee, L. Masinter, M. McCahill: pl. see RFC1357 above.
[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.
[SEAR] Search WWW document full text, http://rbse.jsc.nasa.gov/eichmann/urlsearch.html

[THAU] Thau, R., "SiteIndex Transducer", http://www.ai.mit.edu/tools/site-index.html
[YAHO] http://www.yahoo.com/search.html
[WEBC] WebCrawler, http://www.biotech.washington.edu/WebCrawler/WebQuery.html
[WWWC] World Wide Web Catalog, http://cui_www.unige.ch/