Report of the Metadata Workshop Dublin, OH (1995)

(c) Bipin C. **Desai**
Department of Computer Science
Concordia University
Montreal, H4B 1R6, CANADA
Email: bcdesai@cs.concordia.ca

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 describe and organize information so as to facilitate it's expected discovery and access. With the recent surge in the use of the Internet as a conduit for information dissemination, there is an urgent need to meet the needs of an unprecedented number of information users. Since there would be a significant number of "naive" users amongst them, the need for organizing information and its subsequent discovery must be done with improved functionality and efficiency in terms of 'ease of search', 'time to search' and 'cost to search'. These tasks require the bringing together of professionals from many disciplines of knowledge.

The Metadata Workshop was held between March 1 and March 3, 1995 in Dublin, OH. It was organized by Stuart Weibel(OCLC), Yuri Rubinsky(SoftQuad), Joseph Hardin(NCSA) and Jim Fullton(CNIDR) and 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 (on-line publicly available resources, free or for a fee). The workshop brought together selected professionals from computer science, library science, professional librarians, professionals involved in: on-line information services; abstracting, cataloging and indexing specialists; imaging and geospatial data; museums and archives. The main objective was to address the problem of cataloging network resources with adoption, extensions or modifications of current standards and protocols to facilitate their discovery and access. The list of participants and their affiliations is given in Appendix A.

The goals of the workshop were: to achieve a consensus on a set of core data elements for document-like objects(DLO);to find mapping of these and related elements to accepted standards; and, to provide extensibility to the core set to catalog other types of network objects. It was hoped that the workshop, which was to be 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.

The topics to be discussed were to be the taxonomy of network resources and their characteristics and representation, as well as the established mapping of these characteristics in various cataloging standards. Finally, the identification of a nucleus of data elements to advance standards development and protocols for their use in implementation and evaluation was the expected outcome of the workshop.

Expectations

In spite of the objectives 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 theme 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.

The inter-operability hope may or may not be feasible in an exact manner. For example suppose two catalogs, using different element lists, register the weather information. In one list, the weather is registered by giving average maximum and minimum temperature values for each month while in the other, it is represented by whether the weather is extremely cold, cold, pleasant, hot or torrid. Inter-operability of such systems requires a value judgment which need not be unique.

Nonetheless, there was a feeling that any metadata element set should be properly registered and mapping from one type to another be provided. Since it is never possible to know, in advance, every possible use of any proposed standard element set, it is preferable to create an extensible standard core rather than a comprehensive but frozen one. Any comprehensive standard will quickly outgrow its usefulness due to changes in the real world and ultimately would be suitable for no one.

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 of the user using them 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 semantic content of the 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 if we have no idea about the future. Another concern that was raised was the intended use of the metadata. Different usage may impose different constraints and need different sets of core elements.

There was a concern that the library community's experiences in representing bibliographic information should not be overlooked and be incorporated into the discussions of metadata for networked resources. Issues of authority control, consistency of headings, etc., are critical. Among these were some who were not wanting to reinvent the wheel but were content to adopt a system such as MARC, SGML or TEI etc., and use the selected system or its modification to represent information suitable for discovery.

Some participants realized that in the absence of the general acceptance by nonprofessional cataloger of the already existing cataloging standards (such as MARC which is extremely difficult to understand and use) there is a clear need for an alternate small set of data elements that could be used. Putting into place a simple set of data elements understood by users and providers alike would ensure its use.

Independently developed technologies, such as ARCHIE, GOPHER, WAIS, WWW/http and Z39.50 have been shown to work together, forming complimentary mechanisms for accessing information. It would be profitable if a consistent treatment of metadata across information disciplines could be developed. For example, Z39.50 is now supported by bibliographic and scientific and technical attribute set(STAS). Attribute sets for other disciplines such as business-related data are yet to be developed. By achieving some convergence on metadata expression and usage, general interoperability would be improved, and thus the ability to discover and retrieve useful information enhanced.

Some people, on the other hand were content in coming up with a core element set to enable them to experiment and start implementation. Sets, such as the Semantic Header, were already proposed and used by a number of participants. Their main concern was to see how well their set stood up to the independent scrutiny of the workshop.

Issues

It was recognized that the workshop should try to concentrate on the selection and precise definition of data elements (such as title, author, etc.,) and not concern themselves with their representation (such as Dewey Decimal, MARC, SGML). Issues such as semantics of different annotations should also be addressed. These definitions should be accepted by all communities involved otherwise it would lead to terms used differently

by different communities leading to a Tower of Babel. In this connection the question of 'if and how' to provide authority control was also raised.

Another issue that may become significant is the availability of an object in more than one format (viz. electronic and printed). Should they have the same descriptor or separate ones and the fact that the unique identifiers would not be the same (viz. URL and ISBN)?

Another dimension of the above is the level of granularity of an object. Traditionally components of an object (chapters of books, list of their figures etc.,) are usually not cataloged. Does one use the same metaphor for electronic cataloging? If not how is the relationship of the component to be maintained? This leads to the problem of providing a mechanism to indicate the relationship among various components of the core set and among different instances of the core sets.

How does one deal with a complex object made up of components, each of which can be an object in its own right? The characteristics of the components need not be the same as that of the whole. Handling of revision of the components and its effect on the containing object has to be formalized. How does one resolve/translate attributes that have similar significance? Ordering of components of an element is also significant. The element **name** is a case in point: written in the order <first-name last-name> is not the same as <last-name first-name>. Would this require a meta-metadata which indicates such ordering and semantics?

The natural language used to specify the elements of the metadata and the character set used have to be specified since any default used may not be suitable for all metadata and the objects they document.

There was divided opinion about the nature of the system; whether there was to be a centralized or distributed system to act as the repository of this metadata. If metadata is replicated and distributed, than how does one determine if two incomplete sets of metadata represent the same object? One possible solution is to use a unique name for each object. Another approach is to use the concept of a system generated object ID as used in Object Oriented system.

While it is useful to distinguish the information to be gathered for an object from how it could be used to discover that object, it cannot be done without keeping the users in mind. This requires one to consider the discovery process and how users determine the relevance of an information resource. What are useful elements of any metadata that will aid in establishing this relevance? Any such relevance judgment is of course based on the information needs of the user.

The Workshop: Modus Operandi

The workshop started with a plenary session, during which there were a number of presentations and discussions of the various existing bibliographic standards and proposals The themes of the presentations were: Metadata in Document Management, Digital Libraries and the Web(Larry Masinter); Library Legacies: The Lessons of 30 years of MARC(Priscilla Caplan); The Text Encoding Initiative(M. Sperberg-McQueen); CNI NIDR White Paper(Clifford Lynch); Staying out of the Tarpits: Abstract Element Description(Ralph LeVan). The afternoon was divided into two parts. The first part was devoted to four parallel sub-groups workshops(BLUE, GREEN, RED, YELLOW). The result of these workshops(Appendix B) were reported in the second part to the plenary session in late afternoon. Each sub-group took on the identical tasks of determining the principals and the purpose(use) of the metadata.

There emerged a certain degree of concurrence from the separate workshop subgroup. However, a number of participants who were involved in non-DLO type resources were not satisfied with the results. Others worried about the semantics of the elements, the problem of permanence of the electronic media and that of accessibility and control.

The second day started with the distribution of the printed results of the previous day's workshop. This was

followed by presentations by stakeholders on TEI (Susan Hockey), Indexing and Searching(Bipin C. DESAI), HyTime(Steve Newcombe), CApH(Michel Biezunski), DTD/Support for handicapped: (Yuri), Versions (Barbara Tillet), UR* (Allan Emtage), Bibliography(Lennie Stoval), Addressing(Steve Newcombe). The afternoon parallel workshop sessions of the four sub-work groups dealt with the task of discovery in light of, but not exclusively, the previous day's consensus. The results of these groups (Appendix C) were presented to the plenary session. During this session, each participant noted their preference for features presented in these results. A committee (Terry Allen, Priscilla Caplan, Joseph Hardin, Erik Jul, Daniel LaLiberte, Yuri Rubinsky, Stu Weibel) collated these preferences in the late evening to arrive at a list of core elements and their definitions.

On the third day (a half day that eventually stretched to 3:00pm and beyond), the proposed element sets were hammered out and the work of resolving differences to arrive at Version 1.0 was entrusted to a drafting committee whose membership is given in Appendix D.

Synopsis of Plenary Session Discussions

MARC: Rather than being a monolithic standard, MARC and MARCII exist in formats such as CAN/MARC, InterMARC, OCLC MARC, RLIN MARC, UKMARC, UNIMARC, USMARC(LCMARC), and so on. There is however, no doubt that it is the single most important development in library cataloging and automation, and sharing of cataloging information in the U. S. A. and other countries. MARC is flexible allowing customization and has evolved to encompass newer forms of objects, currently including books, serials, maps, scores, sound recordings, and software. A MARC record of an object gives its detailed description with less emphasize on the supporting objects(source etc.,)

The syntax of the MARC record uses pre-defined tags and data values separated by delimiters. Due to the numeric nature of tags and their 'overloading' even professional librarians and catalogers find it 'hard'. It is unlikely that most 'naive' users would have either the wherewithal or the fortitude to understand, much less provide, information required by the MARC record!

TEI: The TEI is made up of four sub sets: the core, base, additional and auxiliary. It contains an AACR2 compatible bibliographic description of the DLO and its sources. In addition, it contains information regarding encoding, metadata and status of revision. TEI header can be created by the creator of the DLO who may not be a cataloging expert. The header can be used to derive information for cataloging.

SGML etc.,: Using SGML, HyTime, and CApH, it is straightforward to allow metadata to describe itself in a completely generalized and application-neutral fashion, and, at the same time, in a way that is already internationally standardized by the ISO. This means that many kinds of political and economic acceptability problems have already been conquered.

Z39.50 and **STAS**: In the Z39.50 standard, the characteristics of search terms within a query are specified by identifiers called attributes. Retrievable data elements on a given database are referred to as elements. The current Z39.50 BIB Attribute Set (bib-1) contains attributes primarily suitable for bibliographic searching. Since many scientific and technical databases also contain bibliographic data, the BIB Attribute Set supports access to a subset of their data and services. The Scientific and Technical Attribute and element Set (STAS) developed by Chemical Abstract Service(CAS), is an open public definition of scientific and technical data elements.

The objective of STAS is to help support interoperable search and retrieval of scientific and technical databases, using the Z39.50 protocol. However, prior to the development of STAS, there was no standard protocol to refer to a large number of the searchable and retrievable fields within scientific and technical databases. STAS is currently being used in several Z39.50 projects within both the U.S. and Europe. CNIDR is the Maintenance Agency for STAS.

Process of Discovery: A search usually proceeds from one of a general nature to more specific. A search could start with a catalog or review literature and bibliographies. A library user generally searches for objects using the common attributes: subject, title, author and call number in the public access catalogs(PAC). Other search requirements such as words in title etc., may be supported. Details about the object such as the copy-editor, the graphic designer etc., may not be recorded in the bibliographic record, and are not provided as search terms in most PAC. Hence, a requirement to search for books by, for instance, the name of the copy-editor, the head of the team responsible for its printing etc., is not usually provided. Some information of this nature may be recorded in the book itself and is usually difficult to obtain, requiring specialized catalogs or databases.

Document Like Object(DLO): Loosely defined as an object which is similar to an object considered to be a document. Controversy remains as to what it is.

Scope of Core Elements: One of the first issues was to determine the scope of the element set. There were participants who were involved with developing a distributed library for spatial data such as maps, satellite images, digitized images from satellites or other sources. For these participants elements such as author are irrelevant and their concern is with defining geographical locations and features. Other communities, for instance the geological services, require search for and locational details of their collections of geological samples which are not digital, to be accessible with just as much ease as with electronic DLOs. Another example of such a community is a cultural community with a collection of folklore, sound recordings, videos and films.

Versions: The problem of determining versions of a digital document is different than the one involving published work. Two documents which differ from each other by a single bit may not be considered identical. However, such a difference due to an extra space in the text, or a blank line is not considered to be a new edition in the traditional library sense. (Such differences may not occur due to the physical nature of the hard medium). Hence the versioning of a DLO may be determined by the creator(s) of the (intelligence) content of the DLO.

Who will catalog the resource: Since differences such as the one given above may be falsely considered as a new version, the main agent(s)/person(s) to produce the metadata for a DLO should be its creator(s). In addition, the creators usually place their DLO on the Internet and thus act as "publishers". Since the current number of users of the Internet is estimated to be 30,000,000, there is the potential for that many publishers! Since the only reliable authority of the version of a DLO is the creator(s) it is incumbent on them to provide the cataloging information. This could lead to a chaos since usual cataloging conventions may not be followed by all users.

Resolution

In spite of the absence of absolute agreement on all aspects of a proposed core element list, it was agreed to propose one as a strawman version(actually version 1.0, but that number was not unanimous either!) The rationale for this approach was that it was better to do something rather than naught. Even though what we come up initially may not please everyone, we will learn from it while proceeding in a harmonious manner. The alternative was to have nothing as a reference model and have the various metadata designers, and implementors run helter-skelter.

Dublin Metadata Element Set(DMEL) Version: 1.0

Assumptions

1. Common (or Core) element set: These are metadata elements that apply to most/many DLOs.

2. The elements of the Dublin Core are chosen to support resource discovery: finding DLOs and knowing

from them enough about the target objects to know if they will meet the current information requirements.

3. All elements of the metadata are repeatable.

4. All elements are optional.

5. All elements describe the DLO itself except 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.

6. The Dublin 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.

7. Elements are intended to describe DLOs... no attempt is made to assert the suitability of this element set for all possible object types.

8. No assumption is made concerning whether the DLOs are network accessible or specifically electronic.

9. The 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, GILS, TEI, etc.,)

THE PROPOSED ELEMENT LIST

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 which vocabulary.

e.g. English language --- style --- data processing

dogs

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

e.g. Moby Dick: an electronic version

Photograph of the Empire State Building

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

e.g. Melville, Herman

Mao tse-tung(Change in spelling?)

von Neuman Janos

von Neuman, John

otherAgent: the name of any other entity responsible for the content of the object: the role sub-element describes the responsibility(Is there such a need if the author and this is combined with a distinct role subfield)

e.g. otherAgent(Illustrator): Maurice Sendak

<otherAgent role=compiler> John Bear</>

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

e.g. Oxford University Press

OCLC

[Privately distributed]

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

e.g. http://www.cs.concordia.ca/semantic-header.html

(URL)

object-type: conceptual description of object

e.g. map, book, illustration

form: physical, logical, or encoding characteristics

e.g. TIFF ver. 2.3.4.5.6

SGML/TEI P3-1994

35mm microfilm

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). (How about relationship to components etc?)

e.g. Relation(supersedes)(url):

http://laser.cs.concordia.ca

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

Author's Comments

Assumption 3: A specific unique identifier based on a recognized standard is not repeatable, however a given object will often have several unique identifiers each on a specific standard coded explicitly or implicitly. (Dewey, LCCN, ISBN, URN if it ever gets to be a standard, etc.,)

Assumption 4: Since entering the metadata is an option, there is no need to designate all elements as optional. We have to have a minimum set. At least one unique name and/or one unique identifier! We have to enforce some elements to have a value and be eventually part of all DLO's.

Assumption 5: To what extent is this recursion required? One may want to include the source(references, bibliographic list etc., but why should it be recursive? Suppose A uses B as source so we have B as a source in metadata for A. If B uses C as a source, that information would go in the metadata record for B not for A. Recursion is required, in this simple-minded case if C uses A as a source! How can this be possible since C must have occurred earlier in time than B, which occurred before A?Figure1.ps

Should we consider that a different version is a different object?

Even if we ignore versions, the recursion is not in the metadata, but in the system(software) that follows the source links and will not show up in the metadata.

Assumption 6: If the Metadata is to provide information about whether the object it describes is to meet the information requirements, some of the constraints are: cost, the copyright nature of the object, etc., If these are not included, then Assumption 2 doesn't hold!

I think that the user wants the status of the document including its cost etc., up front and hence these should be included (though optional). All documents to be of any permanence should be archived and accessible. Transient DLO may not have a Semantic Header, or what have you but the system that performs the transaction should have one.

Assumption 8: Since there already exist cataloging standards for non-network objects, we should not try to include them here. This initiative is for network-objects which are electronic!

However, they can be discovered and "ordered" via the network.

Assumption 9: If we include in the metadata for an object, the metadata of its components, we could have not only hierarchies but even cycles!

Any complex hierarchy should be specified; however, in my humble opinion, we must not burden the metadata with much complexities. If expert librarians agree that MARC is hard, our metadata will be impossible. I prefer a simpler structure. A simpler structure would eventually be more useful for some 90+% of the needs. Since the object concept already has been developed to describe complex structures, we should look at it. The object approach requires a reference to the components objects. The complexity is absorbed by the managing system not the supplier or user of the metadata.

A structure which asks 'find me the DLO with the quotation: "She who must be obeyed"' requires indexing entire works. Similarly for the structure which describes a mail archive. To find at the outer level whether Joe Public responded to a note from Jill Royal requires a complexity that may not be worth the effort! Such queries should be supported by other means, not by the metadata.

Comments on Element set

Regardless of the language used, the language of the element set as well as the character set should be also coded. This means that we have two sub-sets having as elements say "language and character set". One of these sub-sets is metadata for the element set itself and the other for the DLO. The first is a natural language, the other could be anything including FORTRAN! However, the first is not of concern to the user and should be taken care of by the system.

Subject: Subject element is made up of sub-set made up of two fields, a schema and a hierarchical subject field which includes sub-subject and sub-sub-subject. The classification scheme used (authority) could be specified by using, for each main- level subject entry the schema field entry specifying the cataloging schema. For the hierarchical subject field, the finer level entries must be from the same schema as the top level entry.

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

author, otherAgent: If the order of the name is according to the culture, how is a user, who doesn't know the cultural background of the author to do a search? Also, why not just use the word agent or creator and have a subfield with each indicating the role of the person or non- person(organization) in the resource? What about other subfields such as address etc? Agents of various natures could be coded by using a sub-set, say "ResponsibleAgent", having sub-fields for role and name, address etc., The role/responsibility sub-field could be used to specify the role the corresponding agent has vis-a-vis the DLO. Typical domain of role being: {author, creator, editor, sponsor, analyst, programmer, publisher}.

publisher: Since publisher is covered in ResponsibleAgent, there is no need for a separate entry for publisher.

date: When we are talking about network resources, the date value should not only include the original date of posting (effectively a "publishing date"), but should have a date of modification. This should apply when the DLO is modified rather than changed (say a new version.) Here the responsible agents are the ones who decide on these dates. What constitutes a modification should be left up to the responsible agents. There is no way one can enforce anything except the initial posting date.

identifier: If this is a unique identifier than it requires a sub-set made up two fields to indicate the domain of the identifier and the value. The domain could be a naming authority such as ISBN or LC, and the value could be the corresponding code assigned. For example, ISBN 0-314-66771-7 and LC QA76.9D3D465_1990 represent the same object identifier in different domains. There can be only one value of the identifier for each domain. The problem with the example used above, in the proposed element list, is that we are mixing a somewhat unique name with a location giving a combination which could be repeatable. (e.g., URL made up of "unique" resource name and the location; the same resource could be located in multiple sites with differnt directory paths.)

relation and source: They 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. A book may be made up of a number of volumes or parts details (metadata) of which need to be specified and exist as independent objects. These may be based on another set of objects, etc.

language: As mentioned earlier, 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. Any default used may not be suitable for all metadata and the objects they catalog.

Important elements such as abstract and annotations are **missing**. The former, generated by a human or other agents, gives a capsule idea of the contents of the DLO. The latter is a place-holder for additional details regarding the DLO by responsible agents or others. Without these items the validity of Assumption 2 is questionable.

The boundary of a DLO should be an object and not its component. The components, if significant should have their own element set and be linked, in an object-oriented manner, to the higher level object. The

software should support the liaison. In this way, the complexity of the element set is kept at a manageable level.

Unique name for a DLO should be generated in a way that does not require a central or even distributed name server. Not all objects could be suitable for any DNS based scheme. Each object (or set of identical objects) has characteristics to enable one to create a unique name uniformly. Other identifying characteristics such as the place, time, and/or date may be added to create a unique name that would than become address-independent.

Some control items, not of concern to the end user, must be provided to enforce authenticity, integrity etc.

The element set would have to be stored and searched from distributed and possibly partially or fully replicated catalogs. This requires the cooperation of the sites of the distributed catalogs and of remote software agents. Even though these are beyond the domain of metadata definition, they may have implications on it.

Epilogue

One problem that we recognized was that the workshop spent too much of its time in educating and making presentations which could have been better utilized hammering out the issues at hand.

A working committee, given in Appendix B, was entrusted with refining the workshop element list and reporting this back to the original participants of the Dublin workshop via the listserver. To make this difficult task easier, the listserver was limited to the present participants. However, any comments and concerns of the wider Internet community and those involved in any aspect of the Virtual/Digital Library would have to be channeled via one of the participants. It is likely that there will be another workshop within the next few months to finalize the Dublin Metadata Element (DMEL). Some of the issues that have to be addressed by this sub-committee are: establishing authoritative sources of value for some of these elements(such as AACRII); providing extensibility of the list; and addressing how, if at all the problem of handling of components of the resources would be dealt with.

Appendix A: **Metadata Workshop: Registration List**

Ordered according to registration date/time:

Stu Weibel(OCLC Office of Research),
Erik Jul(OCLC)
Tom Magliery(NCSA)
Kevin Gamiel(CNIDR/MCNC)
Larry Masinter(Xerox Palo Alto Research Center)
Ron Daniel(Los Alamos National Laboratory)
C. M. Sperberg-McQueen(University of Illinois at Chicago)
Diane Vizine-Goetz(OCLC Inc.)
Lynn F.Marko(The University of Michigan)
Steven R. Newcomb(TechnoTeacher Inc.)
Michael Century(Centre for Information Technology Innovation)
Michael Shapiro(NCSA),
Daniel LaLiberte(NCSA)
Joseph Hardin(NCSA)
Alan Emtage(Bunyip Information Systems)
Edward Gaynor(University of Virginia Library)
Susan Hockey(Center for Electronic Texts in the Humanities)
Priscilla Caplan(University of Chicago)
Robert Wolven(Columbia University Libraries)
William E. Moen(Syracuse University)
Roy T. Fielding(Dept. of ICS, UC Irvine)
Dirk Herr-Hoyman(University of Wisconsin-Extension)
Rebecca Guenther(Library of Congress, Network Development)
Ray Denenberg(Library of Congress)
Terry Allen(O'Reilly & Associates)
Yuri Rubinsky(SoftQuad Inc.)
Howard Besser(University of Michigan)
Elizabeth U. Mangan(Library of Congress, Geography and Map Div.)
Bipin C. Desai(Concordia University, Montreal),
Lennie Stovel(Research Libraries Group Inc.)
Michael Mealling(Georgia Institute of Technology)
John A. Kunze(NLM/UC Berkeley)
Dr. Lois Delcambre(Oregon Graduate Institute)
Diane I. Hillmann(Cornell Law Library),
Barbara B. Tillett(Library of Congress)
Andrzej Duda(CNRS-IMAG)
Michel Biezunski(High Text)
Robert E. Kent(University of Arkansas)
Ralph LeVan(OCLC)
Sally H. McCallum(Library of Congress)
Daniel Pitti(UC Berkeley Library)
Pintsang Chang(University of California)
Larry Brandt(National Science Foundation)
Craig A. Summerhill(Coalition for Networked Information),
R. P. C. Rodgers(Lister Hill National Center for Biomedical Communications, US National Library of Medicine)
Avra Michelson(The MITRE Corporation)
Eric J. Miller(OCLC Office of Research)
Les Wibberley(CAS)

Appendix B: Work Sub-group results - Day 1

Compiled by: Bipin C. DESAI

RED Sub-Group SUMMARY

Purpose

Searching

Known item

browsing(fuzzy)

retrieval

display/performance/use

selection

fitness of use

discrimination

management/preservation/achiving

condition of use

format

reference

addressability of component parts

collocation

give "voice" to the objects

BLUE Sub-group Summary

USES

find the object

determine its achival history

search corpora

generate citation

identify the object: author, version

determine operation or applications required

Principles

useful

should support security, authentication, administration of info.

should be able to point to concepts

extensible

metadata attribute to its element?

KISS(Keep it simple stupid)

YELLOW Sub-group Summary

PURPOSE/USES

usable for document

resolve cl?

resolve name to location

build meta database

track provenance and manipulation of data

utility as surrogate

support commercial transactions

current awareness(automatic)

produce citation

produce well-ordered list of base documents

identify source of this version

title if none: what to do e.g. use four bars in music for an untitled composition

versioning of metadata/"duplicate record"

Principles

extensible

scalable- simple to compile

able to describe relationship at different

maintability/up

syntactically independent

no minimum standard of quality or control

version control

documents own quality

convertible

secure from tampering

GREEN Sub-group Summary

PURPOSE/USES

ACCESS: identification, fetches, indirect, forwarding locations, distribution, caching

control: mgmt, integrity, security, version

discovery: searching, browsing, quality, abstraction, distillation

description of data

associated services

Principles

precise, unambiguous

minimal or orthogonal, clean, simple

classifiable elements, typology of element types

easy to use; get extensible to complex use

(scalable)

interoperability, mergeable, distributive

general access, multiple domain, levels of abstraction

fulfill the purpose

Appendix C: Work Sub-group results - Day 2

Red Sub-group Summary

1. content description:

subject - descriptive terms and phrases about the object

example: Metadata

coverage - chronological and/or geographical scope

example: workshop held march 1-3 1995, in Dublin, Ohio

2. identity

title - title

example: oclc/ncsa metadata workshop

version - version

example: Version 7

3. provenance

author/creator - principle creator(s)

example: E. Jul, S. Weibel, Vince ...

publisher - publishing entity

example: OCLC

4. form

genre/type - category of the object (genre); book, article, Home Page

language - natural language of the object: English

system requirements: text/html

5. status

date - date last updated: March 2, 1995, 3:50 EST

conditions of use - limitations governing use or access: unlimited.

Green Sub-group Summary

subject - needs to be structured,

identify the thesaurus and values

title - a value of any data type. attributes include data type of

content,

name of rule set controlling the presentation, structure,

selection, type of title,

author - may repeat. structured field.

name of rule set used.

nature of responsibility of the author.

content descriptor - application specific.(domain specific)

links to/from other docs

keywords

date

identifier

place covered

discipline, object granularity, media type format, thesaurus

language, cost, usage

Blue Sub-group Summary

longer list of elements. allow extend bare tags, but not so detailed.

did some minimal/full examples.

title or name of resource

creator

subject

publisher - of original, digital version. perhaps main term first

unique id- isbn, ssci,

type - book, article, movie,

format - how formatted, what needed to use it

date - of publication, modification, creation, expiration...

copyright statement - qualify with: holder, year, jurisdiction

record source - who wrote the metadata

control number of metadata -

Yellow Sub-group Summary

balance between completeness and conciseness

unique name (URN)

title, at least something descriptive

kind of object (type)

agent (human) author

agent (other)

agent relationship to the document like object (originator, editor, updater,

performer) supply as many agents as appropriate

subject terms.

date of content

date of object - date this derivation was created.

important known relationships; these levels/granularity.

record source - who wrote the metadata

control number of metadata -

Appendix D: Drafting Committee

Terry Allen(terry@ora.com)
Priscilla Caplan(pcaplan@midway.uchicago.edu)
Ron Daniel (rdaniel@acl.lanl.gov)
Bipin C. Desai(bcdesai@cs.concordia.ca)
Alan Emtage (bajan@bunyip.com)
Rebecca Guenther (rgue@loc.gov)
Joseph Hardin(hardin@ncsa.uiuc.edu)
Dan LaLiberte(liberte@ncsa.uiuc.edu)
Tom Magliery (mag@ncsa.uiuc.edu)
Lynn Marko (lynn.f.marko@um.cc.umich.edu)
Sally McCallum (smcc@loc.gov)
Michael Mealling (Michael.Mealling@oit.gatech.edu)
Stuart Weibel (weibel@oclc.org)

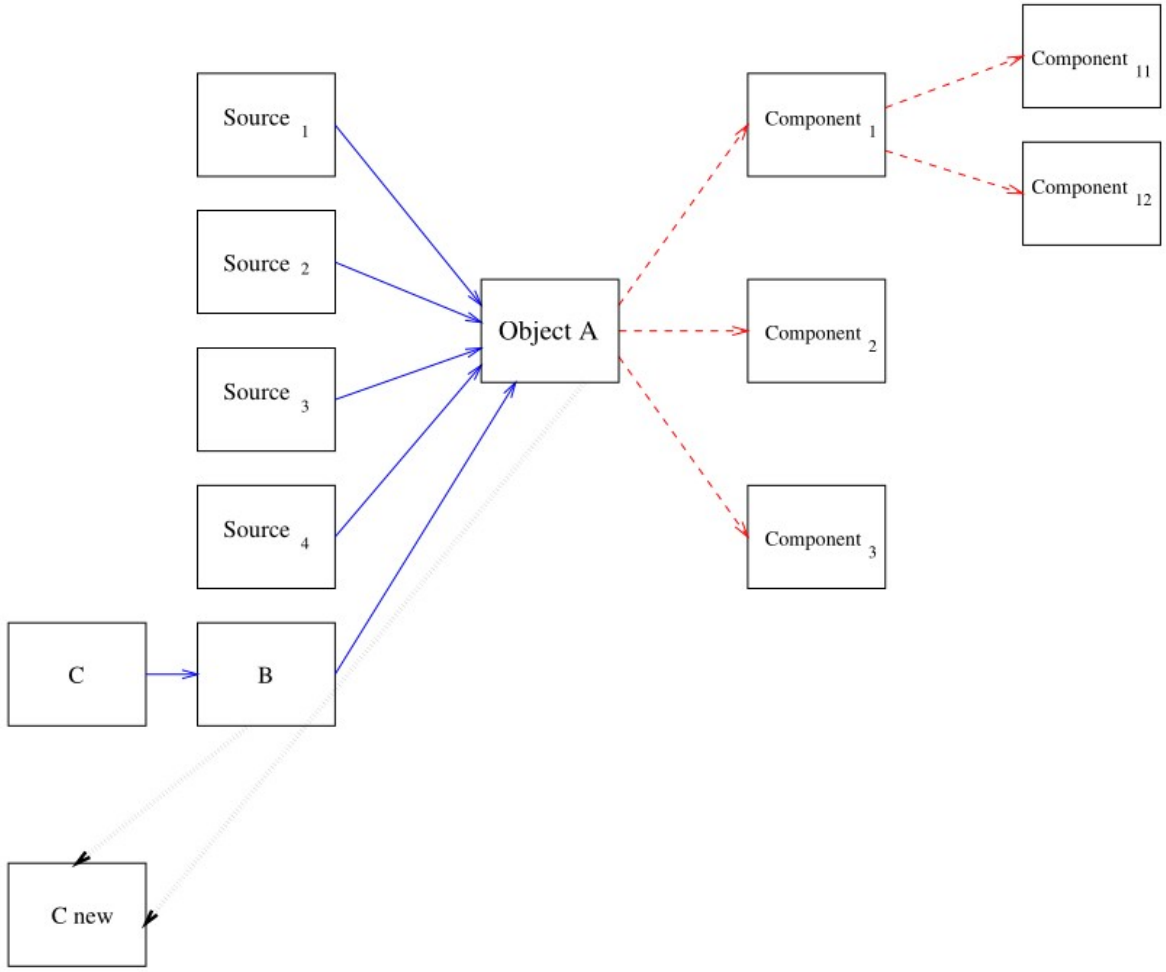

Figure 1: Need for Recursion in Metadata?

BCD