CENDI METADATA INITIATIVES: BEYOND THE BIBLIOGRAPHIC RECORD

Report of the CENDI Metadata Task Group Meeting
December 11, 1997
NASA Center for AeroSpace Information, Linthicum Heights, MD

Sponsored by
The CENDI Metadata Initiatives Task Group
of the
CENDI Cataloging Working Group
and the
CENDI Subject Analysis and Retrieval Working Group

Prepared by
Gail Hodge
Information International Associates, Inc.
Oak Ridge, TN

April 1998

CENDI METADATA INITIATIVES TASK GROUP

Defense Technical Information Center
- Virginia Becks
- John Dickert
- Tanny Franco
- Carrie Schwarten
- Annie Washington

Department of Energy
- R. L. Scott
- Jannean Elliott
- Kathy Waldrop
- Nancy Hardin
- Sue Davis
- Charlene Luther

National Aeronautics and Space Administration
- June Silvester
- Michael Genuardi
- Jacqueline Streeks
- Lynn Heimerl
- Bill von Ofenheim

National Agricultural Library
- John Kane

National Air Intelligence Center
- Dan Karagan
- Bob Steele

National Library of Education
- Keith Stubbs

National Library of Medicine
- Evelyn Bain

US Geological Survey/Biological Resources Division
- Anne Frondorf

CENDI Secretariat: Bonnie Carroll, Gail Hodge

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Commerce, Energy, Education, Defense, Health and Human Services, Interior, and the National Aeronautics and Space Administration (NASA).

CENDI's mission is to help improve the productivity of Federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

"Metadata"(data about data or information about information) has been discussed in many conferences, papers, and workshops within the information and scientific communities over the last several years. The impetus for metadata is that in a networked environment, where the Web becomes an important research tool, it is necessary to describe certain characteristics of the information objects (documents, videos, audio, simulations, images, lessons plans, data sets, etc., that either reside on the net in their full form or can be pointed to via the net). This description aids resource discovery and the use of the information object. The metadata record is intended to serve this purpose.

Traditional bibliographic records provided by the CENDI agencies are also a type of metadata, since they describe (and often act as a surrogate for) the item to which they "point". The bibliographic records have generally been provided by librarians or information specialists using systems and standards that have developed over the history of librarianship. The detailed descriptive nature of these traditional metadata records is difficult to implement in a time of shrinking resources but ever growing information resources. Therefore, new models are being sought that will allow non-professionals to be involved in the cataloging of items that they create.

At the October CENDI Meeting, the CENDI members approved a proposal sponsored by the Cataloging Working Group and the Subject Analysis and Retrieval Working Group to investigate the various metadata initiatives, both the formats and underlying workflows, underway among the CENDI agencies. These initiatives were limited to non-traditional initiatives that either use emerging metadata formats as the basis for restructuring the current bibliographic data elements or that define new elements for new resource types. The aim of the workshop was to inform each other of the various initiatives and to identify common issues and challenges that could be addressed at the CENDI-wide level.

The workshop was held on December 11, 1997 at the NASA Center for AeroSpace Information in Linthicum Heights, MD. Presentations were given by the National Aeronautics and Space Administration (NASA); the National Library of Education (NLE); the National Air Intelligence Center (NAIC); the National Agricultural Library; the Department of Energy's Office of Scientific and Technical Information (OSTI); and the U.S. Geological Survey's Biological Resources Division. The National Agricultural Library was also invited to present its experiences with metadata. Staff from the Defense Technical Information Center (DTIC) and the National Library of Medicine (NLM) participated in the discussion.

The discussions raised a variety of common issues:

- The emerging nature of the various metadata formats makes it difficult to make decisions about their use.

- Emerging formats differ in the degree to which the same information content is segmented or aggregated into the data element fields. This raises issues for validation, searching and data exchange and reuse.

- It is important to be involved in the metadata discussions since the results will have an impact on the development of web-based search tools, on the expectations of our users, and on the degree to which traditional CENDI databases and products will integrate into this new environment.

Recommendations

A single metadata format is unlikely to emerge because of the need to describe different domains and resource types, and to satisfy different user groups. However, key to the ability to discover resources in the networked environment is the development of crosswalks to allow for interoperability. The Metadata Initiatives Working Group recommends that crosswalks be developed between the non-traditional metadata formats described during this meeting. This would avoid the duplication of effort for those agencies contemplating metadata projects in these areas, and promote the development of formats among the agencies that can interoperate (either physically or electronically).

1.0 BACKGROUND

"Metadata" (i.e., data about data or information about information) has been discussed in many conferences, papers, and workshops within the information and scientific communities over the last several years. However, key to its use is the idea that in a networked environment; specifically, the environment of the Internet, where the World Wide Web becomes an important research tool, it is necessary to provide information about information objects that either reside on the net in their full form or can be pointed to via the net. The Web, being primarily a publishing medium, is very text based. Therefore, non-text objects such as videos, audios, images, photographs, computer programs, simulations, and data sets cannot be located in this environment without an accompanying textual description. The metadata record serves both to locate the object and to describe the object to which it points.

It is also the case that bibliographic records have generally been provided by librarians or information specialists using systems and standards that have developed over the history of librarianship. The detailed descriptive nature of these records (as expressed in the standards) are difficult to implement in a time of shrinking resources but ever growing information resources. Therefore, new models are being sought that will allow non-professionals to be involved in the cataloging of objects that they create.

At the October 1997 CENDI Prinicpals Meeting, the CENDI Cataloging Working Group and the CENDI Subject Analysis and Retrieval Working Group proposed an investigation of the various metadata initiatives underway among the CENDI agencies. The proposal was purposely limited to non-traditional initiatives. Therefore, to be considered, the initiatives had to be for a new resource/document type or be attempting to use one of the emerging metadata schemes as the basis for rethinking bibliographic cataloging.

The workshop was held on December 11, 1997 at the NASA Center for AeroSpace Information in Linthicum Heights, MD. Presentations were given by the National Aeronautics and Space Administration (NASA); the National Library of Education (NLE); the National Air Intelligence Center (NAIC); the National Agricultural Library; the Department of Energy's Office of Scientific and Technical Information (OSTI); and the U.S. Geological Survey's Biological Resources Division. There were also participants from the Defense Technical Information Center (DTIC) and the National Library of Medicine (NLM). A list of questions (see Appendix A) was prepared to help the presenters give consistent pictures of the use of metadata in their agencies.

The workshop participants sought not only to inform each other of the various initiatives, but also to identify common issues and challenges. The results of the meeting analyze the proceedings for these kinds of commonalities. Recommendations are also provided for follow-up work.

2.0 AGENCY METADATA INITIATIVES

2.1 Department of Energy, Office of Scientific and Technical Information (DOE OSTI), R. L. Scott, Kathy Waldrop, Jannean Elliott, Sue Davis, Nancy Hardin, and Charlene Luther

OSTI is facing a radical transformation from the traditional online systems to Internet web services. This has expanded the community that would usually access the Energy Database. Database searching by database experts is giving way to easy web access by novices. There is a desire to make database search engines more "user friendly". This is leading to a change in the way that people perceive the need for extensive metadata. Availability of information on the Internet has also changed the view of the comprehensiveness of the database as people move from one database to another more fluidly. The cost of creating and maintaining extensive fields in the database is part of the problem, exacerbated by reduced funding.

Comprehensiveness has been a long-held value and more has always been thought of as better; however, this long-held axiom is changing based on the perceived changes in customer base and data access. It is difficult to tell if this change is right or wrong, but it is occurring. OSTI has already reduced 180 potential data fields down to 137. About 30 are now found in the average record. Proposals being discussed would reduce this number even more.

There is a proliferation of web-accessible database and metadata structure schemes that are merging with search engine capabilities. OSTI is investigating the use of such formats at both the domestic and international level, with the goal to complete the changes to the database structure within one year. There is a Technical Working Group that represents 100 countries on the international database side. OSTI is also working with its Departmental user community. Hopefully, the decisions of both will be very similar in order to ease the processing burden of exchanging records.

OSTI is specifically looking at the Dublin Core initiative as a possible de facto standard for document description on the web. The Dublin Core initiative has gained momentum since 1995. Various workgroups have been formed to work on the details. OSTI decided that, if they were going to streamline to a minimal record, they wanted to be as compatible as possible with something that would be recognized the world over as a standard. They watched the development of the Dublin Core for the first couple of years and had several groups at OSTI evaluating different metadata initiatives. The Nuclear Weapons Information Group was very active in the analysis of various metadata schemes because they wanted something for their program. Various reports were written on the Dublin Core as a result of that activity. Although OSTI realized that they would have to restructure the way the data fit into the fields and also redefine the content of the fields, they decided that Dublin Core would be their guideline.

Currently, they are struggling with what elements should be required versus what are optional. Fourteen of the 16 fields are mandatory. In one sense it looks very easy, but, at DOE, they have so many different record types representing a variety of formats and

carriers; e.g., audiovisual, electronic full-text records, software, etc., that it is going to be difficult to structure an input form that can easily validate the input at the sites, because what is required for one record type will not be required for another. Intelligence will need to be built into the input system, while trying to maintain the simplicity.

The Dublin Core supports subfields and qualifiers (providing a more specific definition of the content of the field or subfield). The nature of many of these components is still being hammered out by various Working Groups of the Dublin Core. In the meantime, OSTI has developed its own qualifier definitions. Dublin Core offers this type of freedom.

The Dublin Core also offers a modular approach. OSTI agrees with the concepts of the Dublin Core that not only must bibliographic data be part of the metadata, but also the administrative data is necessary. Document type, quality type, and certification type data is necessary. In doing the analysis of Dublin Core, qualifiers were identified to handle the administrative nature of the database. There is still a need to manage the record.

OSTI developed its ideas about metadata and has shared them with the DOE Scientific and Technical Information Program (STIP) advisory group. This group includes Headquarters and DOE field staff and representatives from major contractors.

Four STIP goal teams were identified in the STIP strategic plan. One in particular is involved in the development of the minimal record format. The OSTI metadata team did a strawman for the STIP team to review. The group identified mandatory versus optional elements that best fit a bibliographic record. The Metadata team has prepared documentation and the input form that would be required. Concepts for simplifying the input include defaults based on site ID, pop-up menus, etc. An STI stakeholders and STIP meeting is scheduled for January where OSTI will get final buy-in at the department level.

The movement to a minimal record is not as far along at the international level as it is at the domestic level. Early last year, a study was conducted on record simplification within the International Nuclear Information System (INIS) and Energy Technology Data Exchange (ETDE) communities. This dealt with the current record structure, what the users wanted and how important the elements were to the users and database production staff. The consultant also interviewed database producers and publishers to determine what elements were likely to be available electronically from these sources.

This study was presented in April and May, 1997, to the ETDE Executive Committee and the INIS Liaison Officers, who granted permission to conduct a Phase II study. In October, the INIS/ETDE Joint Technical Committee (attended by 26 member countries) decided that the new study would survey the metadata formats available or being developed for, and what metadata would be needed in the Internet environment based on a Web-based scientific culture, rather than working on modifying the current record structure. This study would also include the best output format and relationship of format to products. Some countries also wanted to investigate the use of various metadata formats by primary and secondary publishers. The study will be completed in March 1998. The consultant will investigate applicable metadata formats and supply information back to a team of experts who will review the consultant's findings. The final INIS/ETDE record format will be determined

from the results of this study. Implementation within both INIS and ETDE will be completed by October 1998.

OSTI has prepared three test set crosswalks using the most frequently occurring record types. A fourth test set is a mixed bag of types that occur with less frequency; however, OSTI wanted to make sure the mappings would accommodate these as well as the more common formats. Now that the STIP team has come back with its comments, there will be changes to these.

OSTI distributed a series of tables that map the current EDB fields to the Dublin Core and indicates the types of qualifiers being proposed. They have done numerous mappings but the final decisions will be made when they actually try to dump data into different workflows and document management systems. OSTI is going to redesign the infrastructure internally.

Discussion

Are the qualifier fields DOE specific? Yes, they are, though some of them may also be under discussion by Dublin Core working groups. The international side is still working on their mappings and they may have different qualifiers. Other agencies might have other qualifiers.

Dublin Core is compliant with the Z39.50, "Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnotation" standard. There is a mapping between the Dublin Core elements and the Z39.50 BIB-1 attribute set. DTIC is particularly interested in this as part of the current GILS (Government Information Locator Service. GILS is also Z39.50-compliant and GILS is DTIC's metadata format of choice. DTIC is interested in the way the two would work together. GILS is run primarily by scripts that were created in-house at DTIC. DTIC is now investigating commercial off-the-shelf software for GILS because DTIC has ported all of GILS to the FULCRUM search software to replace the commercial WAIS system. FULCRUM has partnered with Blue Angel. Blue Angel software is Z39.50-compliant and has within it the GILS version 2 and other Z39.50-compliant components. DTIC expects to have an operational GILS prototype based on Blue Angel software in late FY98 with the full implementation complete in FY99.

Chuck McClure recently did an evaluation of GILS. He recommended that Dublin Core be looked at as an option for a replacement for the more extensive GILS meta tags. He also suggested that the validation software created for GILS might be used.

Z39.50 is an important protocol to be considered. Fortunately the Dublin Core Group has very carefully worked with OCLC and that the Z39.50 protocol mapping has been done to the BIB-1 attribute set for Z39.50. In Version 3 of Z39.50, a new set of attributes is being proposed to make the Dublin Core and GILS BIB-1 more similar. There is interest in a smooth merger of the two. Z39.50 has now been accepted as ISO standard 23950, which NISO has accepted.

The question of content rules was discussed. Will OSTI continue to use its current rules? OSTI is in the process of modifying the cataloging rules. Based on ETDE and INIS changes, rules for corporate author input have already been modified. There will be

additional changes during this year as well. These changes will result in issues of products having data in both old and new cataloging formats. There are also use and input system concerns during this transition.

A general discussion of the benefits of aggregate versus segmented (more granular) elements ensued. If you have many elements, then you have the issue of how the search engines will work across elements. If you have fewer elements and you have to develop rules concerning punctuation and order, then it may be more effective to segment. Perhaps the mapping of elements to various record/resource types will help to make the input and quality control more precise even with aggregate elements. There is also the question of how much the input format and the output/product formats should match.

What type of user is DOE customizing the metadata for? There are several user groups: 1) DOE community of researcher and scientists who will be providing more of the input; 2) international; and 3) user groups such as the general public and Federal Depository Libraries. The latter have not been addressed in detail because of time constraints. Novice versus skilled users are being emphasized. OSTI is finding that, on the Web, there are many novice users; therefore, OSTI is taking that into account. There is more emphasis on less complexity because it is assumed or understood that the novice users are not doing the detailed searches that professionals have done in the past.

2.2 National Aeronautics and Space Administration (NASA), Lynn Heimerl and Bill von Ofenheim

As a result of an STI business process re-engineering project, it was determined that the STI program needed to add multi-media products to an array of products and services. NIX (www.nix.nasa.gov) is a new product that leveraged work already underway across the NASA centers. This was an agency-wide cooperative agreement. It was a voluntary effort of STI photographers, librarians, Webmasters, and public affairs staff. NIX started with three centers and by the end had all 10 centers participating. Grants were provided to Dryden, Johnson, Stennis and Ames to help purchase software and servers if necessary. Marshall and Kennedy will develop their own internal databases. NIX was initially released in late May 1997.

NIX points to about 400,000 images, plus Quick Time video and sound. There is a special helpdesk for NIX linked to the NASA CASI helpdesk, which supports the bibliographic products. NIX is a metadata search engine. It is a searchable database that has the look and feel of a central database. It is a distributed database with different search engines. This kept the changes at the centers to a minimum. The NASA Technical Report Server is the model on which this was built. NASA TRS links the technical reports databases across the centers in the same distributed fashion.

Prior to development of NIX three system models were analyzed -- a centralized system, a hybrid system and a distributed system. The centralized system meant that all files had to be stored and searched at a central site resulting in redundancy between the central site and the originating center. This is the most costly model because of the central server development, but it is easier to administer because there is only a single search engine.

The second model is the hybrid, which has centralized metadata records, but distributed image files. Most of the disk space is taken by the images not the metadata, so this is less costly in terms of storage. In this case, you've avoided the worst redundancy. There are several variations on this model with the thumbnails or screen resolutions stored either centrally or distributed.

The third model is totally distributed. All the central server has is the knowledge to contact the distributed sites, collect the information and display it. This is the least costly and least disruptive for the centers. However, the central server must deal with multiple search engines and variations.

The key to the system is a series of PERL CGI scripts that talk to distributed databases. Both HTTP and WAIS (Z39.50) protocols are supported. WAIS was designed for the web and is simple and efficient in the way it performs. HTTP is more generic and often slow.

Searches are performed in parallel. A child process is spawned for each participating center. The child sends the search query to the center and the results are returned. There is a "guaranteed" response time element. A timer is set for each child process's maximum response time. If the child process does not respond, the expired timer is ignored. If a center is down the child will time out and NIX won't be "stuck waiting". The NIX result displays the centers that are not responding.

An unexpected issue arose with regard to the sorting of the result set. Because of the database structure, Johnson Space Center was always at the top of the list when the regular scoring algorithm was used. Instead a "fairness doctrine" has been developed to give all centers equal value on the display. The order is always alphabetical by center. The top ranked item from each center responding is displayed first in alphabetical order by the center, followed by the second ranked in alphabetical order, etc.

The system is easily navigated. The result set brings back a brief description including the title and a thumbnail image. If the user clicks on the thumbnail the screen resolution image is displayed. If the user clicks on more information, the metadata information is displayed. There are different resolutions available and links to the full publications document if available. Each of the center's images or at least the hits are returned.

The primary goal is to make all the distributed databases look and feel the same.

A browse capability is available that uses prebuilt searches optimized for the best quality images on a particular topic. Each one is offered by a different center.

In the future they want to encourage standard key terms in the metadata descriptions and add additional sound, video and Virtual Reality Modeling Language (VRML) objects. They also want to add JPL as another center and the large image collections like the Hubble Space Telescope. There will be an ongoing expansion of the existing collections.

Comments from the users of NIX have been very positive.

Discussion

What software engine is being used? Each of the centers can pick and choose whatever engine they want. Most are using WAIS. Goddard is using FoxPro. Lewes is an SQL database on a MAC. NIX can access them all by changing the submitted query through locally developed PERL scripts to match the input required of the center's search software. Systems administration software is on the central server.

What metadata format is being used? What skill level is required for those entering the metadata, and what kind of thesauri or control lists are being used for the field contents? Metadata is the most critical part of the system. Unfortunately NIX is currently depending on the centers for quality and the content of the metadata. They have asked for specific fields but it is not a requirement. Langley has professional librarians that analyze each of the metadata entries to make sure that the acronyms are expanded and the NASA Thesaurus is being used as the approved source of subject terms. This is the recommendation to the centers, but there is an enormous resource problem.

Is there any technical analysis of the images? Typically, the form is being filled out if the originator wants to include the resource in the Langley repository. Again, this varies from center to center. Automatic image content analysis is obviously needed, but it isn't anywhere near production quality at this point. The form is one created for the photography lab; it is not connected to the Report Documentation Page (RDP) for publications. NASA wants to make the form web based, but it is currently on paper.

How do you make the distributed databases look alike? There are bare minimum requirements. Each database must have a specifically sized thumbnail, specific resolution, metadata, image number and a title. These elements (however they are structured and tagged in the center's database) are then mapped to the NIX format for a more consistent look. The look and feel are well preserved for the users. This is all done automatically by the PERL scripts.

2.3 U.S. Geological Survey/ Biological Resources Division (USGS/BRD), Anne Frondorf

The basic mission of the BRD is to do biological science and to distribute the results of that science. There is an emphasis on information, natural resources and land management.

Part of the original mission of the BRD was to work on creating a national partnership for sharing biological information - the National Biological Information Infrastructure (NBII). This is the biological component of the National Information Infrastructure (NII).

The objectives of the NBII include:

- identification of biological data and information sources in a distributed federation
- easy access, retrieval and integration of data and information from different sources

- application of biological data and information to support resource management decisions

- no unnecessary re-collection of biological data which is actually available from an existing source

In support of these objectives, BRD is not developing a centralized database but, instead, is facilitating the identification and location of the resources of others. They want to aid in making intelligent decisions about the best source, retrieve the information, increasingly combine data across sources, and help researchers and others apply data to their problems and applications

The content of the NBII will be very inclusive, including scientific data sets, information products, lists of experts or organizations, and a variety of software tools to help analyze and apply data to answer questions and share results. BRD is emphasizing the development of the infrastructure needed to support biological information, through the development of standards and common practices and procedures. Where gaps exist, BRD will provide seed money to leverage the resources of others. This is happening particularly in the areas of decision support systems, ecological models, and simulations.

A key aspect of the federation is the idea of a standardized approach to metadata documentation.

The metadata standard needs to be applied to different kinds of content with a biological focus to the metadata so that it makes sense to the user.

BRD began work on a metadata standard based on the FGDC (Federal Geographic Data Committee) geospatial data standard. It is a federal standard for describing geospatial data and is widely used for GIS applications. Since much of the information has a geospatial aspect, BRD was required to use the FGDC. They were also very much involved in the FGDC so there was a commitment to that interagency effort. NSDI (National Spatial Data Infrastructure) and FGDC have been up and running for over three years with a good outreach program and an infrastructure the NBII could leverage. The FGDC extension allows the NBII to broaden into the view of the world that is not specifically biological in nature.

However, the FGDC standard was not envisioned to deal with biological data. FGDC was not set up to let biologists describe their data to other biologists. It is a very detailed metadata content standard with over 200 fields needed to determine precise location and to project to a GIS implementation.

The NBII has a large spatial orientation because of land and resource management, but it also has pure laboratory research that is not spatially addressable. The NBII incorporated the spatial standard and added elements that would be particularly helpful to describe biological data. This resulted in a profile (extension) of the standard that gives a biological view. Particularly relevant is systematics and taxonomic nomenclature - naming of species and higher groups. What naming system was used to name it? Do you have voucher specimens in a museum collection? Answers to these questions reflect on the quality and

relevance of the set for the users needs.

The biological extension resulted in a single metadata content standard for use within the NBII, regardless of whether the data is spatial or not, whether its a data set or information product (publication), etc.

Work on the standard began about two years ago. A working group developed a strawman standard which was then reviewed by an expert panel commissioned through the American Institute of Biological Sciences (AIBS). The profile was then presented to the FGDC standards group. The FGDC approved the proposal to do this, but the spatial metadata standard was just going into a revision cycle. One of the main reasons for the FGDC revision was to establish rules by which different communities could develop the profiles off the standard. The proposal came in a little too soon because formal rules were not yet established. However, the FGDC is almost finished with this revision process. The profile will be redefined as necessary and resubmitted. A full public review process will follow. The international spatial group within ISO is working from the FGDC standard, so there will be a link into an international view as well.

The NBII Metadata Clearinghouse has just been launched. The clearinghouse is a registered NSDI clearinghouse node, so that users may go between the clearinghouses. The clearinghouses are Z39.50 compliant.

BRD has developed a Windows based software tool called Metamaker to allow local input of the NBII standard metadata. This tool has also been adopted by the NSDI. Training sessions have been held for division staff. The session covers what is metadata, why should I do it, and what is the standard, in addition to hands on training on the software. BRD had a naive belief that the scientists would do all their own metadata. Now they are looking at trained and skilled professionals to help with the metadata creation. There is the need for more rigor and control. They are trying a lot of different tacks, including having librarians serve as the focal points for metadata creation.

BRD is also involved in a partnership with NASA and the Global Change Master Directory staff. Metadata from the NBII is contributed to the Global Change Master Directory. The Global Change staff also have a “production” environment and are actually creating metadata for the NBII.

2.4 National Air Intelligence Center (NAIC), Bob Steele

Intelink is a specialized Internet network among members of the intelligence and defense communities. As a closed version of the Internet, it suffers from many of the same problems as the commercial network. It uses web technology to link the consumers and producers of intelligence information. NAIC is responsible for managing the Intelink site. There are Secret and Top Secret Versions of Intelink.

NAIC is typically in the top three in terms of total access with 100,000 – 250,000 accesses per week; DIA is in the 300,000 per week access level. Approximately 50,000 unique IP addresses are currently accessing Intelink. This is misleading because many IP addresses

have multiple users.

The most frequently used products are the finished products that are still done in hard copy. Imagery is another area of interest. Many applications and bodies of knowledge are being developed as a substitute for the hardcopy products. Intelink also provides a variety of tools such as language translation, engagement models, etc. which are also popular.

NAIC is moving away from hardcopy and toward a body of knowledge that represents a specific area from which customized views of that information can be created. NAIC is developing several tools in this area.

ISMIC (Intelink Systems Management Information Center) has oversight of Intelink. They have tried to establish metadata standards with marginal success. There is no real power to enforce the recommendations, so few agencies have adopted them. NAIC decided that metadata would be useful, and they began creating metadata for new submissions about 2-3 years ago. However, they have not standardized as much as they would have liked to. Their users don't want to use search engines - they want "push" technologies ready for them when they turn the machine on.

The problems with attempts to help users find information is that most of the approaches put yet another layer between the user and the ultimate answer. The dilemma is how to help the users find specific answers. Metadata is one of the techniques that can help them do this. It is also necessary to have an overall architecture for the information. Each producing organization is responsible for a different part of the model, but it is difficult to get people to go in the same direction.

Even though there is inconsistent application of metadata, NAIC created some automated processes to aggregate the information -- by country, by the data source and by the product type which groups formal products together, imagery together, etc. Similar aggregation was done around the Intelligence Function Code which tags the subject. Codes have been put on most of the material. The country and subject views have been merged to find the combination of these two aspects.

NAIC decided to develop a standard, even though the intelligence community as a whole did not. Later the intelligence community actually published a standard which they mandated others to use. This was agreed to at the July 1997 Intelink Conference and went into effect on September 30. There are eight required tags, three optional tags, plus nine that are mandatory under certain conditions. The 20 tags include many administrative elements.

NAIC has since modified its local standard to adopt that of the intelligence community. They added to the standard a set of their own tags to make it compatible with Intelink. The prefix for certain elements indicates that they are NAIC specific. There are a total of 30 tags. NAIC is in the process of implementing the standard.

The metadata is created by the analysts as they complete an information product. They fill out a standard form within Framemaker.

There are several uses for the metadata. Most of the users center on making it easier to find information. There is also the need to create alternate views and to perform maintenance on the records. There is an enormous cost for maintenance with over 50 sources on the NAIC site alone. NAIC is also incorporating large databases like the CIRC bibliographic database which has over 10 million document references. There will be thousands and thousands of objects out there. Metadata is critical to managing the site. However, NAIC was not given Intelink committed resources for this; the support is out of NAIC's pocket. Therefore, the more they can automate the process the better off they will be.

NAIC has converted all formal products to the ISMIC standard. Informal products will be completed by the end of 1997. Input systems must be changed to reflect the standard. All views and search facilities will be modified by Feb./March. The main objective was to create software processes to easily change and extend the tags, so they can be flexible and adapt quickly.

Discussion

Are the elements of the standard mapped to CIRC? There was no conscious effort to map to CIRC. Both development teams are aware of what the other is doing. CIRC will be served up on Intelink, but references coming from CIRC will be different. Another issue is how to tag a resource that is being served up automatically from distributed sources. The concept of moving toward an information model and then taking the same data and populating the model automatically would be helpful.

2.5 National Library of Education (NLE), Keith Stubbs

The NLE metadata initiative is the Gateway to Educational Materials (GEM) (/http://geminfo.org). (The Developer's Workbench is at http://geminfo.org/Workbench). The goal is to provide an easy one-stop access to educational materials on the Internet. Nancy Morgan is the GEM Coordinator at ERIC.

At a Spring 1996 meeting of the NLE Advisory Task Force composed of representatives from the constituent groups, the NLE staff demonstrated a collection of lesson plans at the AskERIC Web site. The NLE staff were asked about bi-lingual resources for sixth graders and they were unable to easily locate any using the kind of full-text search capability that is the norm on the Internet. The AskERIC Web page didn't provide real good discovery tools, even though it provided resources.

ERIC contains 30 years worth of references to educational material. The ERIC file is predicated on established channels of publication in print. The web is presenting quite a challenge to that and it moves very quickly. A three-month turnaround to catalog a resource is not good enough. There were also problems with microfiching the variant types of material.

The Federal WebMaster's Consortium involved Mr. Stubbs with the Dublin Core. There was also an effort among the Eisenhower Math Science Clearinghouse submitting resources by teachers, content creators, state educational organizations, etc.

NLE decided that a specialized web resource was a good approach because the common ways of finding information on the Internet --word of mouth, listservs, and own browsing -- are inadequate when looking for educational materials on the Internet. NLE determined that it should sponsor the development of a simple, easy-to-use gateway because the materials are spread geographically, as are the users. No one has a single collection of all educational materials. It needs to have a focus on educational materials that the current search engines do not have at this time. This includes searching by pedagogical technique, grade, subject, curriculum standards, quality ratings, etc. Above all it must be up to date.

Traditional, centralized library cataloging could not keep up. It is too cumbersome and expensive. However, full text search engines retrieve too many hits and frequently mix non-educational materials with educational materials. Also, free text searching cannot distinguish grade level.

A tool was needed somewhere between traditional library catalogs and full text search engines. The following requirements were identified: 1) open standards, 2) distributed cooperative cataloging, 3) adding value to individual collections, 4) customize to appear as a service of that site, 5) each site can choose to point to the master catalog, 6) needs to add value to the individual sites by being done jointly, and 7) a semi-union catalog needs to be able to be produced by the local sites.

NLE has adopted a two-tiered subject vocabulary for GEM. The vocabulary is much smaller than the standard ERIC thesaurus. This is just enough to get you in "the vicinity". More levels could be added if needed. Terms can also be added from the ERIC or NICEM databases, and keywords can also be included as free text. ANSI standards for language and date format have been adopted. The rights management field relates to ownership. Special fields were added to the Dublin Core for educational standards (e.g., fourth grade math standards for the state of Texas), quality indicators, and educational level.

It is a minima list approach, but it is flexible enough to grow so that if more is needed it could be extended.

NLE had an opportunity with NetDay in April to highlight its educational resources. Federal Resources for Educational Excellence (41 agencies total) are trying to tackle this and provide a spot for all the agency resources to be highlighted. NLE is looking at this and other techniques to make the maintenance of this type of site easier. They were able to expand the resource types and the audience (includes parents, students, in addition to teachers).

GEM started in November 1996. The underlying research was conducted by analyzing the questions that people asked of the AskERIC digital reference service. They also asked teachers and performed content analysis of lesson plans available on the Internet. NLE learned that it needed to put the cataloging agency (authority) in the metadata record. NLE has developed a metadata profile and controlled vocabularies. They have also created cataloging and harvesting tools.

GEMCat is the metadata input system (similar to MetaMaker). It is currently available for Windows 95, but Mac is often requested. A Java version may be developed to make the interface platform independent. Controlled vocabularies have been developed for a variety of fields. There is a variety of software to harvest information from various sites and bring it into various subset union catalogs. The testbed of materials is 700+ items and growing weekly. Math Forum. Microsoft Encarta, etc. are being added. Input centers are being considered at Goddard Space Flight Center and the Smithsonian. Two pilot search interfaces have been developed (one on each coast). The one at Syracuse is based on PLWeb by Personal Library Software; the one on the West coast uses Microsoft Access and will be ported to a more robust DBMS.

The GEM project is being done in the context of a number of related initiatives. These projects are intended to encourage the development and collection of high quality lesson plans with a teacher centered review panel. The lesson plans are correlated with curriculum standards and frameworks. The information needed in these metadata records goes beyond descriptive cataloging, but mechanisms for indicating quality have not been well addressed. Related metadata and instructional object efforts (including Instructional Management Systems [IMS], Advanced Distributed Learning Network [ADL], Learning Object Management Group [LOMG], Education Object Economy [EOE], etc.); focused on plugging different pieces of the Internet together to allow for the linkage and control structures. Many of the distance learning efforts of EDUCOM, NIST, Defense Department, and Apple have recently decided to merge together. GEM is in contact with these groups and regularly attends Dublin Core meetings.

IMS is primarily focused on post-secondary education and the defense training community. It is based on large industry training techniques. GEM is largely K-12. However, you can't pidgeon-hole them. IMS is doing metadata based on how you manipulate objects and link them. GEM is focused on describing the objects, not how they interact. IMS wants to have its standard out, but implementing their approach will require new Internet technologies such as XML and RDF. GEM and IMS are developing a memorandum of understanding to ensure future compatibility while preserving the investments that both efforts have made in developing specifications and software. A crosswalk of GEM and IMS metadata is maintained on the GEM web site.

Mandatory elements in the GEM metadata profile include a minimal set of elements (cataloging agency, date cataloged, format, grade levels, identifier, online provider, resource type, subject, title). The identifier for now is the URL. Format and Resource Types have been a confusing issue, since they are still under discussion for the Dublin Core.

The format is the HTML meta tags with additional tags. GEMCat will be able to import/export HTML 3.2, HTML 4.0, XML/RDF, and other formats as necessary. A harvest tool is run to take the records from GEMCat to the site. The ERIC host periodically gathers the files from the distributed sites and fully replaces each site's portion of the union catalog with the updated file. Experimental terms as candidates for controlled vocabulary are collected by the system for human review.

NLE is trying to define a metadata standard that is independent from the U.S. Department of Education. They want to guarantee the harvesting of everything into one place. A clearinghouse is needed to focus on development and innovation. What is being done is not cutting edge production technologies. He hopes that these areas will be improved by others. The National Education Association has already developed a GEMjive, a Javascript alternative to GEMCat.

There remains concern about the ability of the system to scale up. Review and governance structure of the GEM Consortium are also a concern. The quality statement or endorsement by GEM is a major area of discussion. NLE is also interested in developing a lesson and unit submission system for people who don't have an Internet site where they can host their plans.

The GEM Web site has complete documentation for an interchange format and the original metadata profile and controlled vocabularies, as well as conformance files . The NLE search system, Ultraseek, is being customized at ED to recognize Dublin Core and GEM metadata in its search syntax and result ranking.

2.6 National Agricultural Library (NAL), John Kane

NAL has been deeply involved in questions of digital archiving. If an organization takes on the responsibility for archiving, it is also responsible for being able to find archived objects. You don't want to be switching tag sets, etc. in mid-stream. They are working with both judgement and logic. Judgement still requires people to make decisions. Combination of these two factors will make up the system.

NAL has a project to determine how the cataloger's and indexer's environments can be enhanced by computer technology. There was a lot of resistance at first because librarians thought they would no longer be needed, but the aim is not elimination of the librarians but to determine how the librarians roles fit into the unknown of the digital library world.

NAL has been working with SGML as the source of metadata. The DTD is nothing more than the description of the structure of the document. There is much overlap between the SGML and other metadata formats to see how the logical sequence can be brought through from the SGML to support the judgements that the catalogers and indexers must make. Structure adds value, application and clarity. The challenge is to see where we can apply logic and where needed the cost of judgement. People are always more expensive.

A common concern that is voiced about digital archiving is the potential of different native formats that will become unreadable. You can guarantee that in the future people will not be able to read and use them. Therefore, SGML is the more common format. On the Web there are a lot of options in SGML. HTML is an application of SGML (not a very good one). It is difficult to standardize your browser to handle this. NAL, LoC, and several universities have some SGML on the web, but not as much as one would think.

The WWW Consortium communities are reporting back on XML. This has been described as the SGML for the Web. XML will allow non-proprietary structures into documents

which can also carry the metadata. However, the availability and inclusion of metadata will not promise that the full document will be usable by the user in its native format. XML is a valid DTD. This will allow archiving of larger numbers of documents in a more confident manner.

NAL is working with PERL for the structure of the process. The data is in SGML and XML. The structure of the metadata is in Dublin Core and MARC. NAL realizes that there will likely be a number of metadata formats used even within its community. Therefore, they are mapping to GILS, MARC, Dublin Core, AgNIC format, AGRICOLA, and ISIS metadata formats. If new formats arise, they will continue to provide crosswalks to ensure interoperability.

How does NAL enter the metadata? Indexers and catalogers are used. However, they are looking at ways to bring in metadata from non-professionals (via the Dublin Core) and then provide it to the trained professionals for finalizing. The non-professionals could be staff at information centers that do information analysis or professionals that write journal articles.

The current workflow calls for a document to be registered which automatically gives an eight-placed alphanumeric sequence code. NAL is still working in a DOS system, so this could be extended. Indexers give it a number based on the LC heading, but a unique identifier is required. The handle and URN scenarios are likely to be used. NAL wants a sequence of codes that will allow them to work all the way to the digital object. (A record can be as big as necessary. It can hold as many metadata formats as you want. This is a pilot and they want to move to another architecture, because this one may be too slow when scaled up.)

When the document is submitted a Dublin Core input record is created. They have not built in Java scripts that will provide them with more verification and logic. NAL is asking the authors to enter the metadata using their level of skill. NAL uses the CAB Thesaurus. This is copyrighted so they can't put it up on the Net. They are now looking at doing their own thesaurus. They are in the process of building the thesaurus and using it as a filter for the controlled vocabulary. A free text term would be entered in the Dublin Core element which would then be mapped into the controlled vocabulary when it gets to the indexer.

NAL is working to coordinate the information center staff and the needs of the indexers and catalogers. If they get a lot of incorrect information this could be worse than having no initial input at all.

The cataloger responsible for the item is notified that the Dublin Core preliminary record and the digitized hardcopy are available. The cataloger is provided with an SGML document. There is still a question about who completes the GILS fields.

How much will this metadata process cost? Dr. Kane believes that it will be expensive, because it has to be controlled and consistent. However, an Arizona State Study researched how different members of the research community search. Undergraduates looked at the first sentence. Graduate students look at the abstract. Researchers look at the table headings, so they need to go beyond the metadata and look at table captions perhaps within the SGML

document .

There is a great deal of interest in Z39.50, SGML and other technologies for the digital environment. Interoperability and compatibility are important. Blue Angel has merged SGML and Z39.50, but Dr. Kane cautioned that he has found other integration of this kind that is not truly compliant. The MARC DTD has been published. However, it is three megabytes, so LoC is working to modularize.

The National Security Agency has found an Australian company that is compliant with both Z39.50 and HTML (Structured Information Management (SIM) - Kinetic Technologies - http://www.kti.com (Vienna, VA)). Kinetics handles MARC and delivers XML, HTML, RDF or any other format required. Based on the combination of various technologies, Z39.50 may come back as a search technology in a big way.

There is no connection at this time between the system described above and the replacement for the library system. A filter from SGML to MARC will be available by the end of the year.

3.0 ISSUES AND CONCERNS

- Many metadata formats, such as the Dublin Core, do not specify mandatory versus optional elements. This will make it very difficult to reuse data.

- While formats such as the Dublin Core have identified basic elements, there are still questions about inter-element specificity such as qualifiers and sub-elements. Since it isn't known what general practice will be with regard to these structures, it is difficult to make decisions about how to implement such a structure without running the risk of having to make retrospective changes in the future.

- In addition to inter-element issues, the current formats that are being widely discussed may still have elements added for domain-specific purposes. For example, the Dublin Core was modified to add Grade Level for the NLE GEM project. The FGDC geospatial standard was extended to support biological information.

- Validation and quality control, particularly in a distributed, non-professional cataloger environment is a key issue. Many CENDI agencies will have to deal with these metadata records either in their products or through their Web-based search systems. There is concern about the degree to which information will be unretrievable because of lack of understanding and training on the part of the non-professionals who are being called on to catalog this information.

- A key issue for the development of metadata formats is the degree to which elements are segmented versus aggregated. Segmented elements can be more easily validated at input, used in different ways in output products, and can provide field specific searching and sorting. The philosophy of segmented versus aggregated is key to the

discussions currently being held within the Dublin Core community concerning qualifiers and sub-elements.

- It was noted that in some cases centralized "authorities" or clearinghouses were being developed. These entities have some responsibility for quality control, though it may not be at the element content level. Other workflows are beginning to include librarians as the final editors of the metadata records.

- Z39.50 is likely to have an important standards role within the new metadata environment. Additional profiles are being developed for Z39.50. The cultural and demographic subcommittee is about to release its own profile.

- There is discussion about whether metadata formats should be registered. This would provide an indication of the structure that is being used in the record. This may also provide a means for those developing metadata formats to re-use previous work, rather than to duplicate. The same metadata tag should not be used with a different definition.

- The issue of the cost of doing metadata was also discussed. One of the challenges with GILS and various other formats is that they have been unfunded mandates.

- The workshop focused on non-traditional metadata formats. However, there is a need to link to more traditional formats such as MARC in the library systems and the more traditional bibliographic records within the A&I databases.

- The Report Document Page (RDP) can be viewed as a metadata collection mechanism for government technical reports. The question of the connection of the Report Document Page (RDP) with new metadata initiatives was discussed. USGS/BRD has mapped the RDP elements to those used in the metadata scheme, in order to ensure that all elements important for submitting and controlling technical reports particularly to GPO and NTIS have been included. It was also noted that some agencies consider the input from the RDP inadequate as the basis for entry into the traditional A&I metadata records, particularly with regard to the assignment of controlled keyword terms.

- The group also discussed the implementation of vocabularies and thesauri on the web. Lexico has an experimental web-based interface, but there is no evidence on whether people are baffled by this or not. The Community of Science interface has a thesaurus built in to allow the user to peruse elements of the hierarchy in a step-by-step fashion. Compendex and other databases can be browsed. Terms can be selected individually, or the user can switch to a different hierarchy. There is a fair amount of navigation available.
  NASA is looking at using BASISplus Webserver to provide its thesaurus on the web. Multiple frames are available on the initial screen, and an alphabetical perusal can be done. The hierarchical information is presented back to the user. NASA would like to ultimately suggest some natural language expressions related to the controlled term.

- The creation and use of metadata cross a wide variety of communities, including the library, database, scientific, data and museum communities. It is significant that these communities, as well as the standards and technical communities (Z39.50, XML, RDF, MARC, Dublin Core, FGDC, etc.) are communicating. Even though these communities often approach metadata in different ways, as communication increases we can begin to merge the input and the searching concepts better.

4.0 RECOMMENDATIONS FOR FOLLOW-UP

From the review of CENDI metadata initiatives, it is concluded that a single format is unlikely to emerge because of the need to describe different domains and resource types, and to satisfy different user groups. However, key to the ability to discover resources in the networked environment is the development of crosswalks between the formats to allow for interoperability.

The Metadata Initiatives Working Group recommends that crosswalks be developed between the non-traditional metadata formats described during this meeting. This would avoid the duplication of effort for those agencies contemplating similar metadata projects, and promote the development of formats among the agencies that are interoperable

(physically and electronically). Once the crosswalks have been developed, it is possible that a core set of elements for describing any resource type would emerge.