Report on the Second IEEE Metadata Conference (Metadata'97)

Ron Musick*
Lawrence Livermore National Lab
P.O. Box 808, L-561
Livermore, CA 94551
rmusick@llnl.gov

Chris Miller
NOAA - NESDIS - Environmental Info. Services
1315 East-West Highway, Room 15531
Silver Spring, MD 20910
miller@esdim.noaa.gov

1 Introduction

On September 15th and 16th, 1997 the Second IEEE Metadata Conference was held at the National Oceanic and Atmospheric Administration (NOAA) complex in Silver Spring, Maryland. The main objectives of this conference series are to provide a forum to address metadata issues faced by various communities, promote the interchange of ideas on common technologies and standards related to metadata, and facilitate the development and usage of metadata. Metadata'97 met these objectives, drawing about 280 registered attendees from ten different countries and over one hundred different institutions. The audience included scientists, information technology specialists, and librarians from communities as widespread as finance, climatology, and mass storage. The technical program included two keynote addresses, two panel presentations, as well as twenty-three papers and thirteen posters selected from over one hundred abstracts. We provide highlights of the conference below. For more details, the proceedings are available electronically from the conference homepage at: http://www.llnl.gov/liv_comp/metadata/md97.html.

The keynote addresses were "An Architecture for Metadata: The Dublin Core, and why you don't have to like it" by Stuart Weibel, OCLC, and "The Microsoft Repository" by Philip Bernstein, Microsoft.

Weibel's talk described the Dublin core and the Warwick framework - a series of workshops whose output has been a core set of metadata elements common to data from most domains, along with a "container" based mechanism for plugging in larger domain-specific sets of metadata, like the FGDC's standard for geospatial metadata. These efforts represent two of the defining works in this community. Weibel touched on the history of this effort and described his belief that standards such as RDF (resource description framework) for the WWW coming from organizations like the World Wide Web Consortium (W3C) will have a major influence on the metadata community in the near future.

Bernstein's talk covered the Microsoft Object Repository, which also has the potential for a large impact on what metadata gets stored and how they are managed. Bernstein describes the repository as "a place to persist COM objects" (component object model), and as more than just a object-oriented database. The features of a true repository are 1) objects and properties, 2) rich relational semantics, 3) extensibility, and 4) versioning. Repositories are used to help tools interoperate by storing predefined "information models". The information models are the metadata used to describe the underlying COM objects in a standard way such that the objects can be shared across tool boundaries. The main consumers of this type of technology are tool vendors.

2 Catalogs and Interoperability

This session of three talks was chaired by Barbara Bicking of the Environmental Systems Research Institute.

The opening talk by Shklar identified the hurdles that must be overcome to access geospatial data, given the inevitability of a heterogeneous environment. Shklar proposed a federated system comprising a distributed system of catalogs. This system recognizes and accommodates the heterogeneity of

*This work was performed in part under the auspices of the U.S. Department of Energy at LLNL under contract no. W-7405-Eng-48.

different metadata standards (e.g., FGDC and variations or extensions of FGDC). The prototype GeoLens builds and maintains schema standards and conversion tables for the cross-mapping of attributes posed in different schemata, and permits searches and browsing of data based on specification of multiple attributes.

The presentation by Kramer gave a European perspective on the same issue with reference to the European environmental metainformation system Catalogue of Data Sources (CDS). On a national (European) level a core set of attributes is defined, with extensions allowed for specific purposes. CDS draws upon existing international standards like the Global Information Locator System (GILS). A multilingual thesaurus is a key element in the design. As a node in the emerging Global Environmental Locator System (GELOS), CDS shares the same core attributes and has adopted the same standard for distributed search and retrieval, Z39.50. Kramer discussed the status of individual European national catalogue efforts and the prospects for interoperability of CDS with other more specialized catalogues (including earth observation catalogues). An issue that will have to be addressed is interoperability among different keyword groups and thesauri.

Baru described a system being built at the San Diego Supercomputing Center as part of the DARPA-funded Massive Data Analysis Systems project that utilizes metadata catalogs for resource discovery in a distributed, heterogeneous environment of digital libraries. The metadata catalog is divided into four entities: resources (e.g., digital library of images), methods (e.g., library access methods and analysis tools), data sets, and users. The Catalog has been implemented in a RDBMS environment, thus ensuring some portability, e.g., digital libraries that store their metadata as relational tables can be made interoperable. In the future, work on intelligent agents for actively collecting metadata will be pursued.

3 Implemented Models and Systems

This session was chaired by Paul Shelley of the National Information Resource Center (Australia).

In the applications area, Porter addressed issues of research metadata derived from an extensive, long-term field project (the Long-term Ecological Research Sites - LTER). The diversity and spatially distributed nature of the experiments have necessitated the development of metadata exchange standards so as to make the data readily available (e.g., online) and usable. A standard was defined for LTER that has some commonality with the FGDC standard for geospatial metadata and the USGS Biological Resources Division standard.

For the NASA Earth-Observing System Data and Information System (EOSDIS), Klein described the development of a hierarchical metadata structure to support the processing, archiving and distribution of the large volume of satellite, in situ and model data that will populate the EOSDIS Core System (ECS). Interoperability with other metadata systems (e.g., the Global Change Master Directory - GCMD) will be required. Existing metadata standards (FGDC, GCMD) are part of the data model, where appropriate. The data model is made up of 287 elements, but only a subset is needed for basic functionality. Tools have been developed to assist data providers in producing consistent metadata.

Wilkie described the development of a common management structure for the very large, historically heterogeneous, multimedia archive of the British Broadcasting Corporation. This is a major effort to integrate several metadata systems. To the extent possible, elements common to the different media types form the core of the new metadata format. Also defined are the rules for entering information, thus promoting automation of tasks. An indexing language with thesaurus and word association capability has been adopted to focus searches beyond simple keyword searches.

Gillman described an effort at the Census Bureau to build a prototype statistical metadata repository to underpin automated systems for survey processing and information dissemination. The data element registry portion of the metadata model is founded on the the ANSI X3.285 draft standard "The Metamodel for the Management of Shareable Data", which incorporates the principles in the emerging international standard, "Specification and Standardization of Data Elements", ISO/IEC 11179. A major challenge is implementing a process for collection of the metadata. Designing common tools for each (different) survey design and analysis team and motivating designers and analysts to provide the information will require careful planning. A prototype due in October will be able to find metadata across surveys and register metadata objects. The goal is to have a logically centralized but physically distributed metadata repository, which recognizes the desire of the local creators of metadata to locally manage the metadata.

The presentation by Bourdeau discussed the strategy of the Consortium for International Earth Science Information Network (CIESIN) for working with institutions worldwide (libraries, NGOs, government agencies, universities, etc.) to achieve a unified, distributed catalog of data sets managed by each institution. CIESIN deals with information on human interactions with the environment. The original metadata model was NASA's Directory Interchange Format (DIF) but was modified to accommodate the growing diversity of information on the network, i.e., support multiple metadata standards, multilingual capabilities, and tools for metadata development and upgrading. Metadata can be entered into a relational data base through a Web interface in accordance with the major sections of the FGDC Content Standards for Digital Geospatial Metadata (CSDGM). The RDBMS stores metadata in a way that is independent of any content standard. In practice the project has been scoped to support the requirements of the FGDC, DIF, GILS and EOSDIS IMS efforts. For efficiency of search and retrieval, the contents of the RDBMS are exported into text file formats and stored in a Metadata Warehouse.

4 Modeling Techniques

This session was chaired by Matt Morgenstern of Xerox Corporation.

Lagoze described extending the Warwick Framework to expand the resources normally available under the heading of "metadata". He argued that the distinction between metadata and data is artificial and that it is rather the relationship between data sets that is important. This is a novel viewpoint that is certain to generate discussion. Lagoze describes these Distributed Active Relationships (DARs) in the context of digital library repositories. Within this model, metadata packages (that are aggregated in "containers") can be local or remote (e.g., URLs) to the container, virtual or dynamic (e.g., a Dublin Core description could be computed on-the-fly from a MARC record). A prototype architecture (FEDORA) is being developed for a digital library repository that permits aggregation of local and remote content and uses the DARs to control distributions from these aggregations.

Beard defined a meta-information model to open up information resources of a digital library, with an initial focus on geo-referenced data. These resources are represented in terms of knowledge representation systems (KRSs). There are connections of this project to the Alexandria Digital Library project, e.g., a broad perspective on what constitutes a spatial reference framework. A key challenge is the redundancy and ambiguity that arises in assigning names to geographical entities. Extensive testing of the proposed spatial concepts will be required before adequate support of spatial requests can be supported.

A semantic model for hypermedia documents (e.g., interlinked text, pictures, and sounds) was proposed by Froehlich. The modeling language uses the meta modeling formalism of TELOS, a choice based on the viewpoint that semantics defined by logical axioms offers advantages over natural language descriptions. A full-application model suite is defined (domain, navigation, visualization, and user models).

5 Data Warehousing and Integration

This was a five-paper session chaired by Len Seligman of Mitre. The focus was on the use of metadata for integrating heterogeneous data sources. The presenters hit several different aspects of this problem.

Rosenthal spoke first, covering some of the problems the Department of Defense and industry's electronic data interchange are facing in the collection of metadata, the definition and use of intermediate data exchange formats, and the adoption of standards like ISO 11197. He outlined several general approaches for sharing data, and asserted that to connect a large group of highly autonomous individuals, there is little practical choice. One must expect several different interface schema to the data, and hope for explicit correspondences between the schemas. Rosenthal describes a framework for data administration through metadata which 1) describes a standard metadata format, 2) provides proper incentives to use these formats, and 3) unifies the treatment of metadata describing database and data transfer structures.

The use of metadata to help describe external information sources to a centralized data warehouse in a structured, computer-usable way is an important topic that was addressed by two papers. Pu spoke on an approach for using metadata to speed query responsiveness in a data warehousing environment called DIOM. Pu uses metadata to find the set of relevant information sources in an open environment, and determine whether the query is affected by changes in the semantics or description of the component information sources. The main impact of this approach is to reduce the amount of data returned to the user by pruning irrelevant sources and information early on in the query processing. Mazumdar spoke about using metadata to help the user qualify the trustworthiness of various information sources. He described a rule-based system in Datalog for describing the semantic content of the data and constraints on the values being represented.

The last two talks focussed on different aspects of formally modeling and describing metadata and metadata systems. The hope is that if distinct information sources are specified in a manner consistent with the frameworks described herein, then the task of integrating the sources would be much simpler. Morgenstern detailed a formal language for metadata specification called MDS (MetaData Specification). MDS is flexible enough to describe a large variety of data stored in relational, object-oriented, hierarchical and network databases. Kerherve's talk was at a higher level, dealing with conceptual modeling of metadata. She recognized that in a complex information system, there will be metadata at several levels from the system level to the application level. Kerherve plucked apart the different levels, showed a strong correspondence in concepts at each level, and asserted (as did Rosenthal) that an extensible metadata manager must treat all levels in a unified manner to be effective.

6 New Metadata and Novel Approaches

This session on new types and uses of metadata was chaired by Nabil Adams of Rutgers University. The primary focus was on multimedia data and the types of metadata needed to query over such systems.

Zhang's talk layed out the metadata needed to describe video repositories in enough depth for a server to make good decisions on which combinations of repositories to use for different queries. The work includes a framework for video repositories based on having a standard set of image templates and similarity metrics to cluster images. She then defined statistical metadata that describe how the images in any source match up to the templates. This information is very useful in optimizing a visual query.

Shah's talk was complementary, detailing metadata used to describe and cluster images. Shah showed a new scheme for "thumbnailing" by generating a comparison of the image to a null image, and using that as metadata for search. This idea is put together in a clever way that could end up facilitating more advanced distributed image search systems that don't need to know the internals of any participating image retrieval system.

The final paper of the session was presented by Gal. He described a dependency graph that is used as metadata to represent information content on web pages, and then to track how the information content changes. This is proposed as a mechanism to help deal with the constantly changing data sources that we find in highly fluid environments like the WWW.

7 Quality and Limitations

This session was chaired by Kshitij Shah from the University of Georgia. The focus on this session was on the quality of current metadata standards, ideas for measuring that quality, and some limitations of current instantiations of metadata management systems and standards.

Quality of metadata was addressed by two presenters in this session. Moen introduced the results from a significant study of the metadata from 42 government agencies used in the Government Information Locator Service (GILS). GILS is the result of an initiative in the government to identify public information resources available throughout the Federal government, describe that information, and provide tools that help make that information readily available. The study includes a description of the metadata, the criteria that were used to judge it, and a summary of the results. The GILS evaluators identified five dimensions to be addressed in the study: content, technology, standards, policy, and users. This work is an excellent starting point for understanding some of the issues and tradeoffs involved when trying to create a metadata standard meant for broad use.

Conover extended the discussion on metadata quality by speaking to the notion of quality versus quantity. Her thesis was that finding the right metadata is more important than finding the largest quantity of metadata. Their work examined several existing standards for elements that would be useful to the bulk of the users, and applied their ideas to hydrology data.

Daisey's thrust focused instead on the work that has gone on with the Bureau of Census's Common Interchange Format (CIF). In particular, the talk focused on notions of extensibility in metadata and metadata systems. This is an important topic, since the data which are being described in many cases can change enough from year to year or project to project that the deep-level metadata descriptions would need updating and modification. This topic has not yet received much attention in the metadata community.

8 Metadata or Malfeasance: AIIM to Meet the Critical Factors for Success

This panel was chaired by Owen Ambur from the US Fish and Wildlife Service. Panel members included Tom Dale, CADscan, Inc.; Diane Entner, Eastman Software; Ben Kobler, National Aeronautics and Space Administration; Fernando Podio, National Institute of Standards and Technology; and Dan Schneider, Department of Justice. The panel started off with each member reviewing some of the efforts that AIIM is spearheading with regard to metadata standards. Topics included a new proposed standard format for storing data on tape; standards regarding metadata needed to provide document authenticity; a study on the economics of metadata; metadata standards in the document management community; and finally an overview of the legal requirements on government agencies brought on by the Electronic Freedom of Information Act.

The panel got into an enthusiastic discussion on some of the legal requirements being forced on Federal agencies to provide access to all public data. One of the issues is that there are no corresponding funds layed out to help the various agencies comply with the requirements, and so in many cases the work is not moving forward fast enough. Discussion also touched on future directions for GILS, raising some questions about how it will move forward in the future.

9 GILS and FGDC – Dual or Dueling Standards for Content and Retrieval?

This panel was chaired by Steve Hufford from the US Environmental Protection Agency. Panelists included Phil Coombs, Washington State Library; Kristine Kuhlman, University of Maryland Baltimore County; Doug Nebert, US Geological Survey; and Jim Restivo, Blue Angel Technologies, Inc.

This panel discussion compared and contrasted the GILS and FGDC metadata standards, and identified areas where each is of greatest utility. The panelists perspectives varied, since the panel included a state government GILS implementor, two developers of metadata management software tools, and a very knowledgeable member of the FGDC standards community. While the title of the panel implied the possibility of conflict or redundancy between the standards, the consensus of the panel was that both standards have important and different roles to play in helping users find and retrieve relevant information. The complementary nature of the GILS and FGDC metadata standards was highlighted, as was the fact that both are implemented via Z39.50 attribute sets. The great strength of the FGDC standards for completely describing spatially-oriented data resources was emphasized, as was the utility of the GILS metadata format for describing a wider variety of general information resources.

10 Conclusion

Metadata are the key to unlocking the value of the data we are storing. The intent of this conference series is to help build a community within which efficient metadata-based solutions that ensure the identification, exchange and integration of relevant data can be created and shared. These challenges are some of the key limitations that modern data management systems must face as the demands placed on them continue to grow in complexity and size.