

SUL/AIR METADATA TASK FORCE

FINAL REPORT

June 24, 2002

Executive Summary:

Purpose: The Stanford University Libraries / Academic Information Resources (SUL/AIR) Metadata Task Force (MDTF) was charged in early 2002 to provide input upon the descriptive, technical and other administrative metadata elements recommended for use in conjunction with the textual and still image digital resources stored in the SUL/AIR Digital Repository. Attached to this report is the *SUL/AIR Data Dictionary: Descriptive and Administrative Metadata Elements for Digital Text and Still Images, Version 1.0*, and the *Quick View of SUL/AIR Metadata Elements, Version 1.0* which describes the full metadata element set. Also included in the full report is a discussion of other issues related to metadata creation and use at SUL/AIR.

Standards: The SUL/AIR Metadata Element Set is based on existing and/or emerging standards for the various types of metadata including **Dublin Core** for descriptive, traditionally bibliographic metadata, and the Working Draft of **NISO Data Dictionary for Digital Still Images** for technical metadata. Other extensions endorsed to date by the **Metadata Encoding and Transmission Standard (METS)** are still in flux but being considered for other administrative metadata. Metadata structured according to the SUL/AIR Metadata Element Set will be ingested into and distributed from the SUL/AIR Digital Repository using **METS** as the transmission protocol.

Recommendations: Many issues related to metadata were broached in the discussions of the MDTF, but not discussed in depth because of the immediate goals of the group. It is the recommendation of this group, however, that further discussions continue in some manner within the SUL/AIR organizations on issues in the following categories: **information delivery, data modeling and metadata entry, end user documentation, and workflow processes and procedures**. The group would also like to monitor the usage of the metadata element set over time.

Next Steps: MDTF members intend to continue discussions by small group meetings, by establishing a [DigLibWrkGrp Metadata Task Force](#) listserv, and by organizing periodic meetings of the entire group. The most immediate topics for discussion are archival or permanency ratings to collections, data models & workflow (including naming conventions, unique identifiers) based on the metadata element set, end user documentation for the use of the metadata element set and for workflow procedures, and controlled values for some of the metadata elements. Anyone in the SUL/AIR organization or the Stanford University community is invited to join the listserv discussions. Parties interested in joining the listserv should contact either Nancy J. Hoebelheinrich, Metadata Librarian (nhoebel@stanford.edu), or Stuart Snyderman, Digital Library Projects Manager (snyderman@stanford.edu).

Full Report

Introduction

In January, 2002, Catherine Tierney, Associate University Librarian for Technical Services, formed a Metadata Task Force (MDTF) of SUL/AIR staff integrally involved in the development and support of the digital library program. This ad hoc group was asked to (see Appendix A) to:

- Explore the range of uses for metadata
- Define the descriptive and administrative metadata elements required for managing, discovering, and maintaining text and still image formats
- Develop a variety of scenarios for the creation and enhancement of metadata within SUL/AIR
- Identify areas beyond the scope of this charge where SUL/AIR policy is needed.

The following report represents the work of the MDTF, and includes as completed work products, the *SUL/AIR Data Dictionary: Descriptive and Administrative Metadata Elements for Digital Text and Still Images, Version 1.0*, (Appendix B) and the *Quick View of SUL/AIR Metadata Elements, Version 1.0* (Appendix C).

Background Information

Functional uses of metadata based on type:

In discussing the uses of metadata for digital repositories, the MDTF identified three types of metadata: descriptive, administrative, and structural. Each type is important for a particular function related to the digital repository.

Descriptive Metadata

Descriptive metadata is probably the most familiar to library patrons and staff. It typically includes elements to identify who created or authored a resource, its associated date, subject, etc. These data elements are very important for helping end users perform the tasks associated with *finding*, *identifying*, *selecting*, and *obtaining* intellectual resources, regardless of format.¹ The MDTF has determined that not all of the elements in the descriptive metadata set are mandatory.² The minimal level mandatory (M) and mandatory if applicable (MA) descriptive elements of version 1.0 of the SUL/AIR Metadata Element set are mandatory for purposes of *finding* the digital resource (i.e., either title or description, and resource identifier), and of *selecting* the digital resource from among other types of resources (i.e., creator, and resource type). The other descriptive elements (i.e., contributor, publisher, source, date, coverage, subject, language, relation, and rights) are highly recommended because their existence will greatly improve the end user's ability to identify and obtain the digital resource.

¹ The user tasks noted have been described as being the most important by the International Federation of Library Associations and Institutions (IFLA) Study Group on the Functional Requirements for Bibliographic Records. (Chapter 6 User Tasks).

² Those mandatory to the basic functions related to discovery, administration, and structural integrity are identified within the *Data Dictionary* and the *Quick View Of SUL/AIR Metadata Elements*.

Administrative Metadata

Administrative metadata may not be as familiar to the typical library user, but its importance should not be underestimated. Administrative metadata include technical and source metadata, and differ according to the resource type being described, e.g., text, still image, audio, and video. Rights or intellectual property metadata is especially important for preservation of the digital file and for ensuring access to it over time as it is stored in and distributed from the digital repository. Technical and source metadata elements for certain resource types, and for rights, are still being discussed in the digital repository world, and thus, may be subject to revision more quickly than the descriptive metadata elements. Mandatory administrative metadata elements are necessary for *managing* and *preserving* the digital resource over time.

Structural Metadata

Structural metadata describes the relationships between various components of the digital object, i.e., a book to its chapters, a journal issue to its articles, an audio file from an LP to the image of the jacket cover which describes the track, etc. While acknowledging that this kind of metadata is critical to most of the functions associated with discovering, storing and accessing the digital resource, the MDTF recognizes that there are technological solutions available for tracking these relationships using both SUL/AIR's digital asset management system (Artesia's TEAMS) as well as the transmission protocol, METS (METS will be discussed further below). For these reasons, the MDTF neither identified nor defined necessary structural metadata elements.

Objectives in choosing/defining Metadata Element Set

The goal of the MDTF was to provide input upon the choice and definitions of the metadata elements required for managing, discovering, and maintaining our digital collections. Underlying objectives shaping the choices for the elements include the following:

1. To emphasize cross-collection searching:

The SUL/AIR digital library environment contains a number of digital collections/objects that vary in format, extent and type. We expect that the range of collections will extend from complex digital objects containing searchable text images, embedded audio and video files, and interactive executable files, for example, to fully encapsulated course materials that will remain stored until unwrapped for re-use in a new version of a course management system. In the middle of the range, and most prevalent at the moment are digital reproductions of textual materials scanned as images and sometimes marked up as text. From all indications, this range of digital objects currently represented will expand in format, complexity and size.

With the promise of such variety and the increase in capacity, it seemed obvious that a traditional approach to "item level" metadata creation was not realistic. We knew that we could not expect the same depth of analysis for the digital resources that we collect and create as we are accustomed to providing for our physical resources. Such

analysis would be too expensive, slow and cumbersome given the diverse workflow patterns that can be expected for the creation / receipt of digital materials, and the rate at which the materials will appear. Yet, both end users and SUL/AIR staff clearly place very high value on the capability to search across all library materials, regardless of format or collection. Given this expectation, we decided to place the emphasis upon defining a lean, but efficient metadata element set that would allow end users to at least skim across the surface of the many types of resources available, both physical and digital. We expect that more in-depth metadata can be created or utilized for selected collections or digital objects depending upon financial resources, the existence of extent metadata, and the value of the collection to the organization.

2. To facilitate management of a complex digital environment

The diversity of the SUL/AIR digital collections/objects also begs for a simple approach to metadata to reduce the complexity of the tasks associated with the ongoing management of the digital objects. Metadata are going to be not only one of the chief ways for identifying, locating and distributing digital objects, but also for preserving and storing the digital objects over time. The necessity, for example, of mapping existing bibliographic or descriptive metadata in the Unicorn catalog to that of the digital collections will be much more practicable given a simple metadata element set, especially as we were learning how to manage these kinds of resources.

3. To allow global resource exchange

Metadata are also one of the key means of exchanging information about digital content among digital library repositories and / or end users. This capability has been one of the most important features of the MARC element set and record structure even within its seeming limitations for digital resources. The same capability was expected to be equally desirable and useful in the digital repository environment. Currently, most digital repositories in our environment seem to be exchanging metadata using the [Open Archives Initiative Protocol for Metadata Harvesting](#). While decisions have not yet been made to register any of the SUL/AIR digital collections in the OAI Registry, the possibility of doing so needed to be accounted for in defining a metadata element set.

4. To facilitate distributed metadata creation

Judging from the digitization processes used to date by the various parts of the SUL/AIR organization that already have created digital collections, we can expect to see a different model for metadata creation than in the past. While there may be times when professional cataloging staff are available to create metadata for digital collections and to provide training and special consultation on a given digitization project, we expect that metadata creation will be distributed throughout the organization rather than centralized in a couple of locations as with our physical resources. As a result, we expect that those inputting metadata may be less skilled in document analysis than professional cataloging staff. With these expectations, we were compelled to require metadata elements that are relatively easy to use, and simple to understand. Simplicity is also better for cataloging professionals who will

need to catalog digital resources or augment digital records at a high rate of production.

SUL/AIR Metadata Element Set and underlying Standards

The [Dublin Core Metadata Initiative's](#) (DCMI) standard metadata element set was chosen as the basis of the SUL/AIR descriptive metadata element set. It provided the best way for us to achieve the four objectives noted above. With only 15 elements, the unqualified Dublin Core element set should not only be simple for metadata creators and managers, but most importantly, was designed specifically to facilitate cross collection. In addition, is the schema endorsed by the Open Archives Initiative for Metadata Harvesting(OAI) for global resource exchange.

The drawback of using a simple metadata element set is the lack of depth and descriptive specificity for certain types of digital objects such as that of other standards like [MARC21](#), the [Encoded Archival Description](#) or the Federal Geographic Data Committee's [Current Standard for Digital Geospatial Metadata](#). In addition, it cannot be fully mapped to the existing stores of rich metadata in our Unicorn catalog. Fortunately, a protocol is being developed which will allow us to leverage our existing metadata by linking it to appropriate "places" in the digital object. The protocol, the [Metadata Encoding and Transmission Standard](#) (METS) has been developed as a grass roots effort by members of the research library community who have long been involved in creating digital collections. METS works by providing a structural map of the logical parts of a digital object and then either points to or "wraps" the descriptive and administrative metadata associated with those logical parts. If, for example, a MARC record exists for the digital reproduction of a printed book or manuscript, METS can point to that MARC record so that all of the effort of creation and depth of information within it is still available for discovery by the end user. On the other hand, if no MARC record exists, but a SUL/AIR "Dublin Core type" descriptive record has been created, the METS "document" can point to or include that record as it moves both the digital resource and its metadata into the SUL/AIR Digital Repository.

The METS Editorial Board has endorsed various extensions to use for the different types of metadata, including the [NISO Draft Standard for Technical Metadata for Digital Still Images](#). The METS community is in the process of defining extensions to the METS schema for technical and source metadata for text, audio, and visual formats, and for Rights. SUL/AIR will be using the NISO standard for Still Images and has incorporated that element set within Version 1.0. We are also following closely the development of the other extensions and would lean strongly toward adopting them when available, if appropriate for our digital collections.

Scenarios for the creation & enhancement of metadata within SUL/AIR

Due to time constraints and by consensus of the group, specific descriptions of the various scenarios by which metadata could be created and/or enhanced at SUL/AIR were not developed at this time. The group preferred instead to assume that the model for

distributed metadata creation and enhancement would be used as previously discussed. The group chose to leave the task of describing scenarios and developing appropriate workflow procedures for later discussions. See the *Next Steps*, following.

Areas where SUL/AIR Policy & Procedures are needed

During the discussions in the MDTF meetings, a list of issues developed which warranted further discussion in some other type of forum. The list of issues is attached as Appendix D, and includes categories for:

- Information Delivery
- Data Modeling and Metadata Entry
- End User Documentation
- Processes and Procedures
- Element Usage Monitoring Over Time

Next Steps

Members of the MDTF plan to meet periodically to finish up some of the work that the group felt was necessary to complete the group's task as outlined below. In addition, an online listserv has been set up to continue discussions on these and other pertinent issues:

- Establishing criteria for assigning archival or permanency ratings to collections
- Describing and/or modeling data structures & workflow (including naming conventions, unique identifiers) based on the metadata element set:
- Writing end user documentation for the use of the metadata element set and for workflow procedures
- Establishing controlled values for some of the metadata elements such as Creator_Role, Contributor_Role, Digital Publisher_Role, RightsHolder_Type, etc.
- Monitoring the elements' utility over time

Other work products that the group felt would be useful to create include the following:

- A data model based on the final version of the metadata element set and a sample database for metadata creation
- Documentation of recommended workflow scenarios related to the entire lifecycle of the digital objects from material preparation to ingestion into repository
- Draft protocols or procedures for those starting new digitization projects including:
 - what software to use
 - how to handle physical materials, if applicable
 - who should be contacted for consultation of specific issues (such as conservation, media preservation, metadata creation, info delivery, etc.)
 - what the benchmarks or standards are for the format being created, e.g., recommended dpi for archival master TIFFs
- Timetable and process for evaluation of the metadata element set

Conclusion

Establishing the first version of the SUL/AIR Metadata Element Set is an important step in the process of developing a trusted digital repository for SUL/AIR. With the Element Set in hand, digital creators and/or publishers who wish to store their digital collections in SUL/AIR's Digital Repository have much more assurance that their content can be searched, preserved, and managed for the near and long term. As the MDTF has articulated, however, there are a number of other components of the process that still need definition, clarification, and articulation within the organization. While MDTF group members have expressed the interest and desire to continue to work on issues related to metadata, they also have a better understanding of the interrelationship between issues more broadly associated with digital library programs and services. One of the chief benefits gained from the assembly of SUL/AIR staff who participated was the fairly rare opportunity for each to meet and work with colleagues from different parts of the organization specifically on digital issues. Each member brought a variety of skills and areas of expertise to bear on an often confusing array of related topics. Hopefully, the efforts and experiences of this group to work through some difficult conversations and arrive at reasonable solutions will place SUL/AIR in a good position as we enter the next phase of development in volatile digital library environment.

Thanks to members of the Metadata Task Force for all their contributions, including : Paul Zarins, Cathy Aster and Hannah Frost, Julie Sweetkind-Singer, Steve Mandeville-Gamble, Glen Worthey, Adan Griego, Ron Nakao, Michael Olson, Scott Stocker, Christa Easton, Foster Zhang, and Vitus Tang. Thanks also go to other library staff actively in many of the discussions including Walter Henry and Connie Brooks, Media Preservation, and Stuart Snyderman, Digital Library Projects Manager. Questions or comments about this report can be directed to any of the MDTF members or to Nancy J. Hoebelheinrich, Metadata Librarian and Chair.