

UNC Digital Library Project
Final Report
6/30/2000

Table of Contents

EXECUTIVE SUMMARY	4
I. INTRODUCTION	6
What is a “digital library”?	6
Is a digital library necessary?	6
Overview	6
II. FACULTY NEEDS AND DESIRES REGARDING DIGITAL RESOURCES	7
Case #1: Music	7
Case #2: Health Sciences	8
The Digital Library Survey	9
III. A SURVEY OF DIGITAL LIBRARY INITIATIVES	12
Digital library projects at other institutions	14
The William Blake archive	14
Perseus	15
UC Berkeley Digital Library Project	16
American Memory	16
Digital Scriptorium	17
Non-Academic Models	18
The WIRE	19
The Washington Post	19
Digital Library Projects at UNC-CH	20
Documenting the American South	20
MetaLab	21
The Odum Institute / Louis Harris Data Archive	22
FACET (File and Commentary Exchange Tool)	23
The Apollo Project	24
Conclusions and Summary	26
IV A. PROPOSED DIGITAL LIBRARY SPECIFICATIONS	27
Multimedia support	27
Support for common tools	27
Integration with desktop environment	27
Integration with campus systems	28
Resource discovery	28
Data presentation	28
Security	29

Collaboration, user management, and workflow	29
Metadata	30
Why XML?	31
Dublin Core	31
An outline of the proposed system	33
Summary	34
IV B. STANDARDS	35
Images	35
Archival	36
Ephemeral	37
Text	37
Archival	37
Ephemeral	38
Audio	38
Video	38
Metadata.	39
Archival	39
Ephemeral	39
V. IMPLEMENTATION	40
Discussion of capability implementation effort	43
APPENDIX A: COMMITTEES AND STAFF	47
APPENDIX B: DIGITAL LIBRARY GLOSSARY	48
APPENDIX C: UNC DIGITAL LIBRARY DOCUMENT TYPE DEFINITION (DTD)	53
Proposed Metadata Schema	53
Sample XML	58
APPENDIX D: FURTHER INFORMATION	60

Executive Summary

The UNC Digital Library Project was formed to study the issues of digital resource management and delivery at UNC Chapel Hill and to make recommendations on standards for resource management and on the usefulness and feasibility of creating a centralized “digital library” to support teaching and research at the university. The project staff report to a subcommittee of the Faculty Information Technology Advisory Committee. The subcommittee consists of representatives from the faculty of the College of Arts and Sciences, the School of Information and Library Science, and the School of Medicine, and from the staff of the Academic Affairs and Health Sciences Libraries and the Center for Instructional Technology. It is chaired by Steven Weiss (Computer Science) and Bob Henshaw (CIT). Since the end of the Fall 1999 semester, we have also been meeting with the Triangle Digital Libraries Group (TDLG), which is composed of representatives from Duke, NC Central, NC State, UNC, and the Triangle Research Libraries Network (TRLN) to explore the possibilities of a collaborative digital library effort between these institutions.

As a result of our research and discussions with faculty and staff at UNC and other local universities, and with vendors of digital library solutions, we have concluded that it is both feasible and desirable to begin the development of a digital library that supports the range, depth and complexities of the teaching and research missions of the university. These include:

- support for the creation, management and archiving of digital text, image, audio, video, and hybrid resources for individual and departmental teaching and research needs
- robust delivery methods for these resources to permit their presentation, use and reconfiguration across the learning and research environments (classroom, office, library, dorm room, outreach setting)
- centralized support in the form of training, documentation, consulting, systems management and digital storage
- facilities for cross-collection and extra-mural resource discovery and exploitation
- support for collaborative study and analysis at all academic levels (both intra- and extra-mural)

It is the opinion of the committee that there is currently no best-practice example, nor completely satisfactory commercial product, that meets all of these needs simultaneously. If UNC-CH is to implement a digital library that fully supports the institution’s research and teaching needs, it must innovate, emulating or extending relevant features of other digital library projects and adapting and customizing the best commercial products available. The committee’s investigation of vendor solutions has led us to focus on the products developed by Informix (Media360) and Oracle (Oracle *8i* with *interMedia*). The merits of these products are discussed below in Section V . In addition to selecting one of these products, the committee recommends that the following actions be taken:

- Adopt or develop discipline-specific metadata standards for digital resources.
- Work with experts at UNC and elsewhere to pursue a working authentication mechanism to support the delivery of digital resources to users.
- Develop standards and policies for digital copyright management at UNC.
- Continue to explore the possibilities of inter-institutional collaboration.

These needs must be met regardless of the vendor solution that is selected. In addition, both products would require customization in order to meet UNC’s needs. We recommend that the

work of the Digital Library Project continue along these lines, subject to the approval of FITAC and the availability of funding.

I. Introduction

What is a “digital library”?

An informal definition of a digital library is “a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network.”¹ Under this definition, it is clear that digital libraries are ubiquitous. Indeed, there are already many digital libraries in use at UNC. There are a number of different types of digital library in existence that meet this definition. The UNC Digital Library Project is concerned with a specific type of digital library: one that allows for the centralized management and storage of digital resources to support the university’s teaching and research missions. It is this type of digital library that we have been investigating and concerning which we will make recommendations.

Is a digital library necessary?

This is a question that must be answered before we can proceed. The committee believes that the answer is to be found by looking at some of the initiatives already underway at UNC. A number of departments have either implemented, attempted to implement, or are beginning to implement systems to manage digital resources that their faculty wish to use. These systems are very diverse in terms of structure, platform, and support for instruction. They are not necessarily compatible with each other, nor is there necessarily any means for a faculty member in one department to use another department’s system should that be desirable. The lack of interoperability between these systems is a concern, because it creates an unnecessary obstacle to interdisciplinary teaching initiatives. Moreover, since these systems often rely exclusively on departmental resources, there may not be money or personnel available to support them when they degrade or break down. Given that these types of solutions are already being created, it makes good sense at the least to ensure that there are standards in place to aid in their creation and to investigate the feasibility of centralizing support for digital resource management and delivery.

A functioning UNC digital library would provide several benefits beyond the standardization of these diverse departmental systems. The introduction of a uniform interface for resource discovery across disciplines would create new opportunities for the use of interdisciplinary materials. It would allow instructors to focus more on developing content for their courses than on designing and managing systems for delivering digital materials to students. And it would allow technological “have-nots” at UNC (and potentially elsewhere) the opportunity to use digital media without having to divert precious resources towards their development.

Overview

This report will focus on four main areas: the requirements for digital media on the UNC campus, the use of digital libraries both at UNC and elsewhere, the recommended specifications for a UNC digital library, the recommended standards for digital resources, and an assessment of the relative merits of the Informix and Oracle products with an estimate of the effort required to implement either solution.

¹ William Y. Arms, *Digital Libraries*, MIT Press, 2000, p. 2.

II. Faculty needs and desires regarding digital resources

This section will attempt to identify some of the current technology needs of instructors at UNC and to look at ways in which a digital library system could support the university's teaching and research missions. We will begin by examining two cases where faculty and IT support staff at UNC are already taking steps to implement digital resource management systems. These case studies will highlight some of the unique needs of the different disciplines represented, but will also show how those needs overlap. We will also discuss the results of a survey gathered by the project in the Fall of 1999.

Case #1: Music

The faculty of the Music department at UNC wish to develop a system to deliver digitized sound clips to students in their classes. This initiative is a response to the increasing use of digital sound recordings by faculty in their classes, primarily as assignments posted on the Web. The Music department hopes that, ultimately, faculty will be able to create listening assignments by compiling them from existing resources, rather than having to digitize them each time they are needed. The department is initiating the Digitized Recordings Database and Catalog to further this effort. The department plans to pilot the project in the Fall, using MetaLab as the platform to serve the files on the Web. The database will contain the following fields:

- Accession Number (a unique number identifying each recording)
- Composer
- Performer
- Work
- Excerpt or Song
- Identifying Number (Opus or Thematic Catalog Number)
- UNC Library Call Number
- Recording Manufacturer's Label
- Recording Format
- Anthology Number
- Sound File Format

The department faces some interesting problems in the storage and delivery of these recordings:

1. Virtually all of the sound clips to be used will come from CDs cataloged by the Music Library. They are therefore subject to copyright restrictions.
2. The department, in consultation with experts in copyright law at UNC, has determined that the files may be made accessible over the Web under Fair Use guidelines if certain procedures are followed:
 - a) The files must be password-protected, in order to ensure that only authorized members of the UNC community can access them.
 - b) The files must be delivered in a streaming format, so that persons who access them will not be making copies of the files.²

² Streaming formats, such as Real Audio and MP3 deliver their contents to a client as a stream of data, which is played by the client as it is received. With conventional formats, the whole file is copied to the client and then played.

3. Faculty in the department are reluctant to spend a great deal of time in providing cataloging information for the digitized files, especially since they derive from CDs already cataloged by the Library.

Given these requirements, the department must pay great attention to the security of the digital resources they use, and must, at the same time, provide a system that is convenient for faculty to use.

Case #2: Health Sciences

The School of Medicine at UNC is currently engaged in a project to develop a database to index digital resources for use in instruction. In some respects, the needs of this kind of project differ substantially from those of the Music department, but as we will see, there is also substantial overlap in some of their metadata needs. The schema for the new database is still under development, but the list of fields will likely include the following:

- title
- creator
- subject, subject vocabulary
- description
- date resource was created
- format, format scheme, format resolution, file size
- identifier, identifier scheme (the unique identifier for the item, and its location)
- coverage
- rights information

The resources to be indexed will primarily be images at the outset, but text and other type of resource may also be included. Some of the issues peculiar to health sciences are:

1. In some disciplines, e.g. radiology, very high resolution images in lossless formats are the minimal standard for usefulness. There is thus a need for massive amounts of storage space and high bandwidth for the delivery of such images (if they are to be stored centrally).
2. The need for patient confidentiality necessitates strict control over access to certain resources.
3. For some purposes, there will be a need to capture additional metadata, e.g. information about the patient, such as age and medical history, and diagnostic information.
4. For teaching / training purposes it will often be necessary to present resources to students with only certain subsets of the metadata available, in order to allow them to make their own diagnoses, for example.

As we can see, there are requirements that span both cases, even though these may result from different imperatives. The faculty in both cases need to be able to store and reuse digital resources in a way that both maximizes convenience and at the same time protects those resources from unauthorized use. The most significant requirement of all is the one driving both projects: it is becoming increasingly clear to faculty and IT support staff at UNC that well-organized systems to manage digital resources are indispensable. There is simply too much effort and time being spent on discovering, creating, and managing digital resources by individual faculty, teaching assistants, and staff—effort which is frequently duplicated and which

runs the risk of taking time away from the development of course content. In effect, without effective resource management systems, technology can be an additional burden on instructors.³

The Digital Library Survey

As part of the UNC Digital Library Project's mission, we conducted an online survey of instructors and staff at UNC which attempted to identify the basic requirements for a digital library to support teaching and research. There were 119 respondents to the survey, of whom 73 are faculty, 29 staff, and 17 students. As might be expected in an online survey, a high percentage of the respondents are already incorporating digital media into their teaching, but a much higher percentage want to do more than they have been. Some of the questions and a breakdown of the responses are listed below.

How/where do/did you use digital media?

- 56.3 % Teaching in the classroom
- 53.8 % Collection management (e.g. using a database, creating or maintaining a Web site, etc.).
- 45.4 % Research
- 41.2 % Teaching outside the classroom (e.g. homework assignments)
- 34.5 % Collection development (e.g. scanning slides, creating digital audio, etc.)

How would you like to be able to use digital media?

- 73.9 % Collection management (e.g. using a database, creating or maintaining a Web site, etc.).
- 72.3 % Teaching in the classroom
- 67.2 % Research
- 62.2 % Teaching outside the classroom (e.g. homework assignments)
- 61.3 % Collection development (e.g. scanning slides, creating digital audio, etc.)

From the two responses above, we can see that, even though a sizable percentage of respondents are already employing digital resources, more would like to do so. The fact that the issue of resource management ranks highest in the list of desired uses of digital media reflects the difficulty presented by this task. It is worth noting that the differential between answers flattens out considerably when compare current usage to desired uses, from a spread of 21.8 percentage points to 12.6.

What functionality would you like to see included in a UNC digital library?

- 76.5 % Availability of digital materials contributed to the collection by other members of the university for your use.
- 75.6 % Sophisticated searching and data mining tools for finding useful materials in the campus collection.
- 74.8 % Facility for adding your own materials to a collection shared with colleagues within the university.
- 66.4 % Automatic generation of Web pages from materials you select from a multimedia collection, including imagery, audio, texts, etc.

³ A 1998-1999 survey by the Higher Education Research Institute at UCLA found that "Information technology is the fourth most frequently cited source of stress among women (74 percent) and the fifth most frequently cited among men (64 percent), ranking for both groups above other sources of stress such as research and publishing demands (50 percent), teaching load (62 percent), and the review and promotion process (46 percent)." See <http://www.gseis.ucla.edu/heri/faculty.htm#stress> for details.

- 65.5 % Ability to limit access to digital materials or collections (e.g. to a particular class, or group of colleagues).
- 55.5 % Automatic generation of digital course packs from a multimedia collection.
- 49.6 % Facility for assigning comments, keywords, or other information to a particular object in an online collection, whether or not the object belongs to you.
- 48.7 % Facility for adding your own materials to a collection shared with colleagues outside of the university.

Some other responses:

- pan-campus work-flow
- facility for linking from materials to dedicated discussion group/newsgroup
- development of geolibraries
- collaboration with digital libraries at other universities/institutions
- ability to track different versions of digital assets; intellectual property rights tracking (e.g. digital watermarks)
- help in clarifying copyright issues

As we can see from the responses to this question, there is a good deal of support from our respondents for most of the potential functions of a UNC digital library. The possibility which elicits the most enthusiasm is the idea of being able to share digital resources with the rest of the university community. The creation and management of digital resources is a time-consuming process, and one which the instructors who responded to our survey would like to be able to make as efficient as possible.

In addition to asking about desired functionality, we also wanted to know what kinds of institutional support users would like to have. There was broad support among our respondents for institutional help with the issues of storage, management, and delivery of digital content.

What support would you like to have the university offer for such a library?

- 70.6 % Authoring tools to facilitate creation and management of objects in a digital library.
- 64.7 % Digitization of existing collections on campus.
- 64.7 % Support and maintenance for presentation of digital materials in the classroom.
- 63.9 % Support and maintenance of central facilities for storing your digital materials.
- 62.2 % Acquisition and licensing of commercial digital collections.
- 60.5 % Training (e.g., data entry, content retrieval, etc.).

We also asked for free-form responses regarding any special requirements that members of the UNC community might have for digital media. A selection of these responses is listed below.

Do you have any special requirements for the quality, resolution, or manipulability of digital materials that you use or would like to use?

- Help in getting materials entered into existing (or new) databases, because it is so time consuming.
- I have a large slide set from years in countries of Asia that I would like to digitize and select for class, homework, coursepack, assignments, and other possibilities. They need their color and especially focus as much as possible. I would also like to be able to edit short, say three minute, sections from a VHS video and use them, sometimes combined, to illustrate points in a lecture instead of needing more than half an hour to work through the video.
- Foreign language texts (Arabic, Turkish scripts) integrated with English (Latin script).

- My digital data is large. One set requires about 200mb. I currently have about 20 sets, and plan on getting many more. I need some room to store these for other people to have access to them.
- Quality should be very high. Currently, slides remain a better quality product than digital images--perhaps this will change. DUAL PROJECTION is a must. Viewing of single images is not satisfactory. Without the ability to use images simultaneously, most art historians would find this database useless.
- Since I have only limited knowledge of the potential here, I don't know what to expect. Certainly, I need remedial training in adapting my old 2x2 slide driven lectures into a format that students can access from a web site.
- I have some fairly large (over 100 GB) data sets. These are hard to manage on departmental or ATN servers, and will be growing. Multi-terabyte scale disk farms would be a welcome addition to UNC's resources.
- Sound quality adequate for foreign language learning, with precise random access and replay of sound segments.
- Speed of access is important. The files are fairly large (20MB+). We'd like to share the data between campus units (CS, Physics, Materials Science, Chemistry and Gene Therapy so far) and off-campus units (NIEHS in RTP).
- High frame rates for video. Optimal compression rates for video and audio Virtual Reality and modeling, especially for human movement. Greater use of QuickTime standards for enabling technologies in video and audio for cross-platform ease of use. Digital video decks for use by faculty/students, not controlled by a multimedia office, including disk storage on-site, not controlled by another office across campus---give the controls and tools and storage to departments and programs--they can learn this and many of us want to, without university structures and limitations on access!!!
- Sound files (MP3 etc) that are of high enough audio quality that they can be used for music analysis and research.

One theme running through the responses is the need for high-quality digital resources. It is frequently the case that digital materials (especially on the Web) are optimized for delivery to users with relatively low bandwidth (e.g. 56k modems). While such materials are useful, they are usually not of high enough quality to be useful for university-level teaching or research. The recommendations for minimal standards in Section IV of this document address this concern, but there must also be storage space for the larger, high-quality files. The campus' participation in Internet II may ultimately address the growing need for more bandwidth, but the issue of delivery will continue to be a difficult one as long as there are potential users accessing digital content through dialup connections. The solution is to provide ways of accessing either Web-optimized or high-quality versions of digital media, depending on the context which is why we recommend creating both high-quality archival versions of digital resources and also smaller surrogate files.

III. A survey of digital library initiatives

Three ingredients will be essential to the success of any institution-wide digital library initiative at UNC: emulation of successful strategies and capabilities found in other digital library systems, avoidance of shortcomings identified in those systems, and accommodation of the features, needs and limitations of existing digital library systems already operating within the university. For this reason, the UNC Digital Library Project has conducted an extensive survey of digital library initiatives both at UNC and elsewhere. In this section, some of the initiatives investigated by the project are presented, in order to highlight particularly salient “best-practice” examples, as well as significant shortcomings or limitations that should be avoided in the design and implementation of UNC’s digital library.

The identification and evaluation of these “advantages” and “caveats” has been governed by attention to three major aspects of digital library systems in general:

1. All digital libraries fall somewhere on the spectrum between **resource-focused** and **collection-focused**. Resource-focused libraries tend not to make strong distinctions between the digital resources they house. In other words, a more resource-focused library will probably not store two items from different collections in two different places, with two different types of cataloging information. Collection-focused libraries are, as their name implies, comprised of one or more discrete collections of material, which may have completely different characteristics and capabilities. The distinction is rarely a definitive one: a digital library might store all of its resources in a single database, but deliver them to users as members of different collections, for example. Or a digital library might store all of its resources in different places according to the collection of material to which they belong, but provide ways of searching across those collections.
2. Also important is the degree of **interconnectedness** in a digital library. This is a measure of the extent to which relationships between discrete resources are created. For example, given a particular resource, how many actions does a user have to perform in order to move to a similar or related resource? Can he simply follow a hyperlink from within the first resource to the related resource(s), or must he initiate a search for the related object(s)?
3. **Mission** constitutes the third major aspect of all digital library systems. A good digital library system is created to meet a clearly-defined set of institutional needs and to serve a clearly-defined group of users. The most successful digital libraries are those that implement features and provide resources that are dictated by clearly-articulated mission and a sharply-focused understanding of users’ needs and interests.

The components and capabilities of any digital library system are further influenced by a number of factors, including: types of content, financial and technical resources available for creation and maintenance, administrative and institutional policies. The table below outlines some of the related issues that have been considered in summarizing the systems described in this section:

Storage	Management	Delivery
<p>formatting</p> <ul style="list-style-type: none"> • standardization • support for multiple formats <p>archiving</p> <ul style="list-style-type: none"> • requirements / reasons • space / media • formats • maintenance 	<p>metadata</p> <ul style="list-style-type: none"> • catalog records facilitate search • catalog records facilitate user decisions in selecting a resource • creating catalog records requires time / expertise <p>rights</p> <ul style="list-style-type: none"> • security • intellectual property <p>history</p> <ul style="list-style-type: none"> • history records facilitate change auditing and content creation • history records facilitate error recovery <p>collections</p> <ul style="list-style-type: none"> • single vs. multiple • discrete vs. interconnected 	<p>search capabilities</p> <ul style="list-style-type: none"> • single vs. multiple ways of accessing the data (access points) • options / effectiveness depends upon robustness of metadata <p>browsing</p> <ul style="list-style-type: none"> • pre-defined vs. user-configurable arrangement of contents <p>user interaction</p> <ul style="list-style-type: none"> • user-specified display of content? <p>accessibility</p> <ul style="list-style-type: none"> • function and ease-of-use for discovery and delivery of content to users of differing mental and physical abilities

These issues fall into three categories: storage, management, and delivery. In the storage category, we have considered issues like :

- What is to be stored and in what format(s)?
- Are the objects being stored intended to be a permanent or long-lived resource?
- How much space is required to store archival-quality versions of the objects?
- Is there a plan for the long-term maintenance of the library?

Second, in any digital library, there should be ways to manage both resource discovery aids and resource life cycle. Critical questions include:

- How are the resources in the library going to be cataloged, and by whom?
- Is there a mechanism for tracking how resources were created and / or modified?
- How are the resources organized? Are they stored as separate collections? Are there relationships between individual resources?

Finally, how are the resources in the library delivered? The options for resource delivery depend on the other two categories in many ways. For example, if a group of resources is stored as static Web pages, then there is likely to be only one (or at most two) ways for a user to access those resources: by following the hyperlinks that the developer of the Web site has created, or perhaps by using a search engine that has indexed the site. If the resources in the library have useful metadata attached to them, then searching for resources is more likely to be easy and productive. If not, then resource discovery may be difficult or even impossible. If resources have been stored in or indexed by a database, then both searching and browsing may be more productive,

depending on how the library's content delivery tools have been set up. Some further considerations are:

- How many paths does the library provide to its resources? Possibilities include searching, browsing through a system of hierarchical menus, browsing through interlinked resources.
- Is the user allowed to decide how content is displayed? Can she specify what information is to be displayed in a search return, for example?
- How accessible is the library? Can it be used by people with visual impairments, for example?

With these considerations in mind, let us examine first some digital library projects at other institutions, and then see what has been done here at UNC. For each project, we provide a summary of that project's mission and focus, and an assessment of its capabilities and limitations as they apply to UNC's needs.

Digital library projects at other institutions

The William Blake archive

<http://jefferson.village.virginia.edu/blake/>

Mission

“...the Blake Archive was conceived as an international public resource that would provide unified access to major works of visual and literary art that are highly disparate, widely dispersed, and more and more often severely restricted as a result of their value, rarity, and extreme fragility.”⁴

Focus

Preservation and presentation of the works, in multiple media, of a single individual; presentation and relation to relevant scholarly materials (e.g., bibliographies).

Assessment of features

Narrow (by design) but robust and innovative, tailored to the nature of the content and the perceived interests and needs of the user community. The project is well-supported, and has stayed true to the mission and vision it began with.

Summary (Blake):

Advantages:

- Clear, well-defined mission
- Features implement mission effectively
- Standards use and archiving exemplary

Caveats:

- Narrow, topical focus dictates practices and features different from those required for a multi-disciplinary institutional digital library

⁴ <http://www.iath.virginia.edu/blake/public/about/glance/>.

Perseus

<http://www.perseus.tufts.edu>

Mission

“The Perseus Project is an evolving digital library of resources for the study of the ancient world and beyond. Collaborators initially formed the project to construct a large, heterogeneous collection of materials, textual and visual, on the Archaic and Classical Greek world ... Recent expansion into Latin texts and tools and Renaissance materials has served to add more coverage within Perseus and has prompted the project to explore new ways of presenting complex resources for electronic publication. Our primary goal is to bring a wide range of source materials to as large an audience as possible. We anticipate that greater accessibility to the sources for the study of the humanities will strengthen the quality of questions, lead to new avenues of research, and connect more people through the connection of ideas.”⁵

Focus

Creating and facilitating static and user-configurable interconnections between and within the components of large, heterogeneous collections of text, image and geographic data resources all tied to the general rubric “study of the ancient world and beyond.”

Assessment of features

Features are varied and well-conceived in support of the project’s mission and focus (inter-connectedness). Research and discovery are facilitated by on-the-fly and pre-processed creation of context-sensitive links between resources without requiring the user to explicitly request them or conduct a search. Most features are robustly built using open-source tools, but the effectiveness of some suffer from implementation bugs or non-intuitive user interfaces, difficulties in part incurred by the need to build sophisticated functionality from scratch (a condition imposed by virtue of the selection of development and production environments). Despite such minor difficulties, Perseus has achieved deserved acceptance within the academic community as an important and unique resource for teaching and learning about the ancient world.

Summary (Perseus):

Advantages:

- Focus on interconnectedness of resources highlights capabilities some UNC users will certainly require

Caveats:

- Bugs and idiosyncrasies result from large scope and “build-from-scratch” approach; better to acquire critical indexing features through a mature commercial product?

⁵ <http://www.perseus.tufts.edu/PerseusInfo.html>.

UC Berkeley Digital Library Project

<http://elib.cs.berkeley.edu/>

Mission

“Re-inventing scholarly information dissemination and use. The UC Berkeley Digital Library Project is developing the tools and technologies to support highly improved models of the ‘scholarly information life cycle.’ Our goal is to facilitate the move from the current centralized, discrete publishing model, to a distributed, continuous, and self-publishing model, while still preserving the best aspects of the current model such as peer review.”⁶

Focus:

Development of tools and technologies. The Berkeley DLP is very much a test bed and futurist think tank for a particular class of scholarly activities, rather than an effort to build a coherent, functional digital library.

Assessment of features

Innovative approaches to tools and media formats (from image searching to geographic cataloging and collaborative document authoring) hold great promise for improving and enriching the features that future digital libraries may be able to provide their academic users. Much of the innovation is tied tightly to the proposed paradigm shift in “scholarly information life cycle” articulated in the project’s mission statement. Collections associated with the Berkeley DLP have been assembled to facilitate development and testing of candidate technologies; their arrangement and articulation is therefore often wedded to the particular development question or need under investigation.

American Memory

<http://memory.loc.gov/>

Mission

“The National Digital Library Program is an effort to digitize and deliver electronically the distinctive, historical Americana holdings at the Library of Congress, including photographs, manuscripts, rare books, maps, recorded sound, and moving pictures. To achieve its goal, this unique public-private program, also works in cooperation with members of the Digital Library Federation and other libraries and archives throughout the United

Summary (Berkeley):

Advantages:

- Demonstrates that tools, technologies and formats will continue to evolve, changing the requirements and opportunities for digital libraries and their user communities: a UNC digital library must be modular and scalable to accommodate such innovations

Caveats:

- Test bed supports a narrow research and development mission: not a good model for a permanent digital library system supporting the research and teaching missions of a major university

Summary (American Memory):

Advantages:

- Multipath browsing and robust cross-collection searching a “best practice” example worthy of emulation

Caveats:

- Failure to provide easy, persistent bookmarking of resources vitiates much of the value of the resource: the path from resource discovery to use in context (e.g., the classroom) must be quick, easy and reliable).

⁶ <http://elib.cs.berkeley.edu/>.

States...The American Memory Historical Collections [is] a major component of the Library's National Digital Library Program...The program will augment ever-growing services provided by traditional libraries. Online primary sources held by many different institutions will be made available through the National Digital Library Program. By increasing access to these sources, it will enhance the broad intellectual and research support already provided in libraries and classrooms."⁷

Focus

Facilitating discovery (via searching and collection-browsing) and access for users of all ages and abilities to materials in a variety of media (text, image, audio, video, hybrid), served from multiple locations, and all united by their identification as "historical Americana."

Assessment of features

The American Memory / National Digital Library Program provides access to its constituent materials via both a topically-organized collection finder / Yahoo-style directory (permitting discovery, then browsing or searching of individual collections) and a user-configurable, cross-collection search engine. Users may construct searches against any subset of the collections, and limit searches by such aspects as key word and media type. Emphasis has been placed upon a uniform user interface, clear and complete representation of cataloging and copyright information, and ease of navigation (one always knows which collection one is in, and what its general subject is). Capture of found resources (by bookmarking or URL copying) is greatly hampered by the fact that URLs returned in search results are relative to that particular search, and therefore persist only as long as that search result set is active (only a few minutes). Documentation provides guidance on how to capture the "permanent" URL for a given resource, but the method described therein is complicated and requires both patience and significant technical facility on the part of the user.

Digital Scriptorium

<http://SCRIPTORIUM.LIB.DUKE.EDU>

Mission

"...to support the Rare Book, Manuscript, and Special Collections Library's mission of providing access to historical documentation through the use of innovative technology and collaborative development projects with Duke University faculty, students, and staff..." This mission is embodied in the three major goals of the scriptorium:

1. "To enhance access and aid preservation"

By providing scholars with digital versions of library materials that are too rare or fragile to be allowed to circulate, as well as tools to browse, search, and analyze these materials remotely via the Internet, they will be able to do their research more quickly and from a location that may be more convenient to them.

Summary (Digital Scriptorium):

Advantages:

- Robust, standards-based cataloging of resources.
- Multiple paths to resources via browsing or searching.

Caveats:

- Reliance on a tool (DynaWeb) that is no longer supported by its maker (INSO Corp). This problem is mitigated by the project's standards-based implementation, which will facilitate the migration to a new system.

⁷ <http://memory.loc.gov/ammem/helpdesk/amfaq.html>.

Broader access to these digital facsimiles reduces the need for handling the originals and aids in their preservation for future scholars.

2. “To add value”

Library resources are enhanced and new discoveries facilitated by providing scholars with rich background and contextual information about the materials as well as sophisticated text and image manipulation and analysis tools to which they might not otherwise have access.

3. “To facilitate research, teaching, and learning”

The Digital Scriptorium aims to use advanced technology to bring more researchers into the library through collaborative projects with faculty and students, thereby encouraging them to come to the library to use resources in innovative ways, and to leave the fruits of their scholarship with the library for future researchers.⁸

Focus

The Digital Scriptorium is the extension into cyberspace of Duke University’s Rare Book, Manuscript, and Special Collections Library. It is focused on individual collections within the library, but does provide some cross-collection searching capabilities. The digital collections on the Scriptorium site are developed and cataloged by a permanent staff of librarians on a collection by collection basis as funding for such development is acquired.

Assessment of Features

The Scriptorium’s handling of metadata for their digital collections is exemplary. A number of their collections are marked up using the Encoded Archival Description in Standard Generalized Markup Language (SGML). This format allows for validity checking and a high level of consistency in their metadata.⁹ The project’s attention to standards and best practices in the creation of their collections makes them less dependent on a proprietary delivery system for their content. Even though the software they use to deliver information to the Web (DynaWeb, from the INSO corporation) is no longer supported by the company that developed it, their use of an open standard to catalog their resources means it will be possible to migrate to a new system in the future with a minimum of inconvenience. The truth of this statement is evident in the ease with which some Digital Scriptorium collections have been incorporated into the “American Memory” project (q.v.) and into the UNC digital library prototype.

Non-Academic Models

To round out our survey of external digital libraries, we should briefly consider how they are used at institutions that have missions other than education. Virtually all large companies (and many smaller ones) employ more or less centralized methods to manage their digital resources. We will look at two organizations that use digital libraries to achieve missions similar to the one we envision at UNC. Both manage and archive digital information and provide tools to their employees to aid in structuring and delivering that information as needed. These

⁸ <http://scriptorium.lib.duke.edu/scriptorium/about.html>.

⁹ See <http://scriptorium.lib.duke.edu/findaids/ead/> for information on the Digital Scriptorium collections that use EAD and for links to EAD and SGML resources.

organizations, the Associated Press's *WIRE* Web site, and the Washington Post, employ technologies which the committee has researched and evaluated for this report.

As news organizations, both companies are, like instructors in an academic setting, in the business of information delivery. Both have realized significant savings in time and effort¹⁰ by investing in the development of their own digital libraries. Both firms have developed a database-driven digital information management system that facilitates the entire digital media workflow: from resource creation and discovery, to selection, modification and delivery, to archiving and reuse. In this sense, both of these commercial digital libraries are better examples than the non-UNC academic digital libraries discussed above for the kind of broad-topic, multi-media digital library system envisioned for UNC in this report. In most of the academic digital libraries, content creation and life-cycle management is carried out behind the scenes by a staff of technology professionals, rather than by the content experts (i.e., academicians) themselves, who may serve only as project directors, advisors or simple users. In these commercial examples, an information management system puts the subject-matter experts back in control of content generation and management, with the technology experts serving in a support and advisory capacity.

The WIRE

<http://wire.ap.org/>

“The Associated Press (AP) has more people in more places covering more news stories than any other news operation—from the heartland of the United States to the heart of Africa. As the world’s oldest and largest news-gathering organization, AP is an innovator in news transmission—and has been from the days of the telegraph. Working together with Informix Enterprise Consulting Services, AP is using Informix Dynamic Server™ with Universal Data Option™, and Web Integration Option™, and the Excalibur TextSearch DataBlade module to deliver a 24-hour continuously updated multimedia news site. The site, called *The WIRE*, is linked and customized for more than 230 affiliated sites operated by AP member newspapers and broadcasters.”¹¹

The Washington Post

<http://www.washingtonpost.com/>

“Every day, the [Washington Post] creates hundreds of news stories—including photos and graphics—that must be archived and made readily available to journalists working on deadline. At the same time, the Post’s valuable information assets must be easily repackaged for online publishing...The Washington Post uses Artesia Technologies’ TEAMS, together with a

Summary (commercial):

Advantages:

- Demonstrate value of off-the-shelf products in developing an author/archive-to-web solution
- Robust back office / work flow components put subject-matter experts in direct control of authoring and management, not technical gurus

Caveats:

- Off-the-shelf products do not provide the full suite of enterprise-specific capabilities; local customization and extension required--true also for UNC
- Media and rights needs are similar to UNC's, but mission and user community (and therefore features, capabilities and even architecture) differ significantly

¹⁰ Perhaps money also, although neither of the articles referenced below contains a cost-benefit analysis.

¹¹ <http://www.informix.com/informix/success/aspress/aspress.htm>.

powerful Oracle database and Oracle ConText (now part of Oracle *interMedia*), to organize and leverage its news content across multiple publishing channels.”¹²

Digital Library Projects at UNC-CH

Documenting the American South

<http://metalab.unc.edu/docsouth/dasmain.html>

Mission

“Documenting the American South (DAS), an electronic collection sponsored by the Academic Affairs Library at the University of North Carolina at Chapel Hill, provides access to digitized primary materials that offer Southern perspectives on American history and culture. It supplies teachers, students, and researchers at every educational

level with a wide array of titles they can use for reference, studying, teaching, and research.

Currently, *DAS* includes five digitization projects: slave narratives, first-person narratives, Southern literature, Confederate imprints, and materials related to the church in the black community. Another project featuring North Caroliniana is in development.”¹³

Focus

DAS is a collection-focused project, which, like the Perseus Project, deals with a relatively narrow subject area. Like both the Digital Scriptorium and Perseus, DAS employs SGML to mark up the electronic texts that comprise its collections. To date, the project has produced an impressive array of digital texts. The texts are marked up in Text Encoding Initiative (TEI) conformant SGML and are then transformed to HTML for delivery to users who do not have access to an SGML browser.

Assessment of Features

DAS is a tremendously important initiative, providing public access to primary sources for secondary, undergraduate and graduate teaching and research that are available nowhere else on the web. The aggregate of system design, standards-based implementation and staff expertise that comprises DAS makes it UNC’s “best-practice” example for the digital encoding, archiving and delivery of textual materials. DAS sets the example for other projects at UNC that intend to digitize, web-publish or archive significant textual materials for teaching and research use. DAS could benefit from a robust UNC digital library system that would facilitate multiple search and browse options now being implemented by projects like that at Berkeley and some commercial media sites, as well as “interconnectedness” features such as those found in digital libraries like Perseus. Page-by-page searching and viewing, “live” links from words and phrases in texts to multimedia objects housed in other collections, and automatic indexing and concordance tools for textual analysis are all value-added features that a robust digital library system could bring to DAS’s content.

Summary (DAS):

Brings to a UNC digital library:

- Methods / expertise for digital text encoding
- Unique, high-value content for use in UNC curricula and outreach

Benefits from a UNC digital library:

- Multiple search and presentation options (automatic / on-the-fly)
- Interconnectedness with other digital resources

¹²

http://success.oracle.com/success_demo-wwwprd-dcd/plsql/partner_display.show?i_customer_id=426&i_story_id=1416.

¹³ <http://metalab.unc.edu/docsouth/aboutdas.html>.

MetaLab

Mission

“The MetaLab Project at UNC-CH is the primary site in a network of information services provided by key universities around the world. MetaLab operates as a library, a publishing house, a distribution center and a technology showcase. The materials available on MetaLab represent a diverse community of information providers who obtain space on MetaLab only after meeting the following criteria:

- Do the materials further the teaching, research, or public service mission of UNC?
- Does the collection use technology in innovative and unique ways? Every collection need not be innovative, but it should use up-to-date technology.
- Does the collection add synergetic value to other MetaLab collections? Does it complement or contradict other collections of music, agriculture, politics, religion, software, etc? An answer of NO to this question should not necessarily disqualify a collection; we may want to begin a new collection area.
- Is all of the material copyright clear and otherwise legal? Exceptions for “fair use” may apply.
- Can and will the keepers of the collection operate in a self-sufficient manner or provide requisite support funding? Exceptions can be made for especially important collections.
- Is the collection non-commercial or operated by a not-for-profit organization or individual?
- Are the materials of national or international interest? Individual pages are permitted for those who are contributing to other broader collections, but parochial materials should be kept on the campus web server.”¹⁴

Focus

MetaLab is an excellent example of digital library that combines the features of a collection-focused project (like Duke’s Digital Scriptorium) with those of an information technology test bed (like Berkeley’s Digital Library Project). As articulated in MetaLab’s FAQ, MetaLab is “a collection of information technology experiments funded by a variety of sources: some academic, some corporate, and some small information technology start-ups.”¹⁵

Assessment of features

MetaLab houses many collections, presentations, data stores and virtual communities with topics or emphases ranging from African American Literature to metal working to political action to Linux to Yiddish language and culture, each with its own diverse array of content types,

Summary (MetaLab):

Brings to a UNC digital library:

- Broadest experience in digital collections management and digital media on campus
- Unique, high-value content for use in UNC curricula and outreach

Benefits from a UNC digital library:

- Standard, user-reconfigurable interface options and cross-collection, meta-data driven searching
- Interconnectedness with other digital resources
- More tools for creating and managing diverse media content

¹⁴ <http://metalab.unc.edu/collection.html>.

¹⁵ <http://metalab.unc.edu/metafaq.shtml#3>.

discovery/delivery methodologies and content presentation features. Each collection or project hosted on MetaLab is designed to preserve access to, or facilitate the distribution of, important content and innovative presentation of broad public interest or lasting public value. Each is therefore either rehosted (with or without modification) from another site, or designed and implemented as a stand-alone project by MetaLab staff and the content originators. Some of these collections are static, or even no longer updated by their originators, but kept “alive” on the web as a public service because of their continuing value; others are tremendously active, serving broad user constituencies both on- and off-campus.

MetaLab and its staff constitute the most diverse and active source in the university community for examples of, and expertise concerning, digital media, presentation, standards, and the like. MetaLab’s content is of inestimable value for teaching, research and outreach purposes. The effective design and implementation of a UNC-wide digital library system could benefit significantly from the creation of a constructive relationship between MetaLab and the Digital Library Project.

One important aspect of the digital library concept developed by the UNC Digital Library Project has been an architecture which would facilitate the continued operation and independent management of important digital initiatives like MetaLab while facilitating cross-collection resource discovery, dynamic “interconnectedness” with items in other collections, and a user-driven discovery-to-annotation-to-presentation workflow. Such an architecture will enable the UNC digital library system to add value to MetaLab by providing it with new tools and additional technologies to enhance, expand and preserve its own content under its own editorial rules while providing standards-driven bridges between sub-collection level objects within MetaLab and other campus and extra-mural digital collections in the context of various learning and research environments. Designed in cooperation with MetaLab, a digital library system could, for example, provide on-the-fly, customizable keyword linking between a DAS text, civil-war era songsheets in Duke’s Digital Scriptorium, audio clips from the Southern Folklife Collection, and images from the Library of Congress’ American Memory collection, all without requiring students or instructors to code a single line of HTML.

The Odum Institute / Louis Harris Data Archive

Mission

The Odum Institute maintains the oldest and the third largest archive of machine-readable, social science data in the U.S. Its Louis Harris Data Center is the exclusive national repository for Louis Harris public opinion data. The Institute also maintains an extensive collection of U.S. Census data, including one of the most complete holdings for 1970 Census files, as well as other types of national and international economic, electoral, demographic, financial, health, and public opinion data. The Odum Institute also provides permanent, standards-based archiving for social science data generated by UNC researchers in order to make it available to other researchers in the future.

Summary (Odum Institute):

Brings to a UNC digital library:

- Broadest experience in data archiving and management of large statistical datasets

Benefits from a UNC digital library:

- New venues, tools and users for its data holdings

Assessment of features

Data warehousing is another important function for digital libraries. Several disciplines require the use of very large data sets, and the Louis Harris Data Archive is one of the few initiatives on campus which supports this need. The Odum Institute maintains and expands its archive holdings in context of a broad range of training, statistical analysis, grant assistance, making its staff the premier on-campus source of know-how for the acquisition, standard cataloging, storage and maintenance of large datasets like census and opinion poll results, as well as geospatial information. Just as cooperation with MetaLab would bring a campus-wide digital library valuable expertise and content in the area of digital media, so a synergistic relationship with the Odum Institute's data archiving activities would place high-value datasets at the disposal of a broader array of institutional users while simultaneously insuring that digital library plans for archiving and handling of large datasets is influenced by experienced, knowledgeable advisors.

FACET (File and Commentary Exchange Tool)

Mission

“FACET, Phase I, is a small application that helps students, teachers and staff organize themselves into private groups, then share messages, documents and web pages with other participants in their own groups. Group participants can create subgroups containing all or just a few of the original group's participants.”¹⁶

Assessment of features

FACET is an entirely different type of digital library from those we have considered so far, but it meets the definition in that it stores and manages digital resources and delivers them to users on demand. It differs also in its structure: unlike most of the libraries surveyed above, FACET requires that users install client software on their computers. This client application talks to a database which stores the documents to be shared and manages the users and groups that have access to those documents. FACET represents an interesting application of digital library technology to fulfill an educational mission. Its functionality is independent of subject area, so that it can serve the needs of any instructor who needs to share and annotate documents within a group.

Experience and methodology gleaned from the implementation and use of FACET is critical for the development of the “user side” of a UNC-wide digital library system. As discussed in detail in section V, the unique enterprise needs of a major research and teaching institution like UNC will necessitate significant customization and extension of whatever commercial products are purchased in order to implement the digital library. Of particular importance in this context are the paths between digital content authoring / management and the learning environment, whether the conventional classroom, a coffee shop on Franklin Street or a distance learning situation (like a public library or community computer center in rural Cherokee county). Sharing and annotation of documents, and the ability

Summary (FACET):

Brings to a UNC digital library:

- Model for digital sharing of commentary and annotation (students and instructors)

Benefits from a UNC digital library:

- Rich range of media to annotate and comment

¹⁶ http://uncled.oit.unc.edu/Facet_1/overview.html.

to easily relate those documents or annotations to materials an instructor (or student) has selected for discussion or analysis will be critical to the adoption and success of a UNC digital library system. FACET provides an important model for the implementation of such features on a broader scale.

The Apollo Project¹⁷

Mission

The Apollo Project comprises an online database of digital media objects related to the study of the civilizations of the ancient Mediterranean and medieval worlds together with WYSIWYG tools for the discovery, selection, annotation and presentation of those objects in the context of UNC undergraduate and graduate courses.

Focus

Like Perseus or DAS, the Apollo project's content is limited by association with a particular set of themes, in this case "the study of the civilizations of the ancient Mediterranean and medieval worlds." In terms of "interconnectedness," Apollo occupies a middle ground, for each digital object is cataloged separately in the database, without subordination to a particular thematic collection. These objects acquire temporary membership in one or more "virtual collections" through the actions of instructors, who use a GUI-driven interface to select, annotate and group them for presentation or use in their courses as web-based slide shows or thumbnailed "media sets" (essentially, digital coursepacks which instructors and students can access 24x7 from the department's web server using only a standard web browser). These media sets can be associated with individual class dates in the semester calendar.

Apollo's permanent archival records employ the standard descriptive element set developed by the Visual Resources Association--the same cataloging scheme employed by the Art Department's slide library, from which some of Apollo's content was digitized. The contents of each descriptive record are supplied by an individual faculty member or graduate assistant when the associated object is first added to the collection (a prerogative of any instructor using the system). There is no systematic oversight or standardization of the cataloging process. Individual instructors can subsequently supply their own annotations and, in a limited fashion, control the visual presentation of materials as they select them for use in a given media set.

The overall focus of the Apollo Project is on the organization and delivery of digital media in the learning environment, as supported by the archiving and reuse of digital objects as well as a delivery mechanism that exploits web technologies without requiring users to author web pages directly.

Assessment of features

The Apollo system has been used by a growing number of instructors in the departments of Classics and History since the fall semester of 1998,

Summary (Apollo):

Brings to a UNC digital library:

- Exemplary model for multi-mode delivery in learning environments / academic processes
- Minimal technology skills required of novice users: shifts focus to content and context

Benefits from a UNC digital library:

- More robust policies and support for metadata creation and standardization

¹⁷ The Apollo Project was largely developed by two members of the Digital Library staff, Hugh Cayless and Noel Fiser.

during which time a significant quantity of data and a number of feature enhancements have been added to the system. During this two-year period, over 1300 students in some 20 undergraduate courses and seminars have benefited from the increased exposure to course-related digital media (in particular, imagery, documents and maps that would otherwise have constituted only a limited aspect of the course experience). Most first-time Apollo instructors have continued to use the system in subsequent courses, and in at least one case it was an undergraduate student who, having seen Apollo used in a previous course, suggested that her instructor employ the system to distribute course handouts so as to save paper and copying costs. Instructor testimonial indicate that there are two primary reasons for the successful adoption and continuing use of the Apollo system:

- the course- and date-focused delivery system, which brings digital media into the classroom and homework environments without imposing a significant technology skills burden on instructors or students, and
- the ability to put materials before students for an entire semester that previously could only be addressed fleetingly in the course of a single lecture, so that deeper, more analytic questions can be asked about them and more sophisticated and contemplative responses expected

Apollo's success, and its causes, illustrate the importance for the Digital Library Project of customizing and extending best-in-class commercial products to meet the unique enterprise needs dictated by UNC's teaching and research missions. Early adoption and frequent reuse by faculty, instructors and students will depend not only upon the range and robustness of the system's features for media authoring, management and discovery, but also upon the ease and effectiveness with which users can employ the system in the delivery of its contents as part of the various processes (e.g., course design, lecture, homework, paper research, conference presentation) which constitute the day-to-day academic business of the university.

The Apollo Project's complete dependence on its individual users to supply metadata is one potential weakness that also deserves attention during the design of a campus-wide digital library. Although the Apollo Project implements a standards-based cataloging scheme in its software and interface, as we saw in Section II, faculty are often (understandably) unwilling to devote the time required to generate truly robust metadata. In addition, users cataloging images scanned from slides tend to rely on the data printed on the slide itself, if any, which may be inaccurate or incomplete. In a campus-wide digital library system, materials deriving from established archives and contributing projects can be expected to exhibit some degree of uniformity and standards compliance with respect to descriptive schemas and nomenclature; however, some significant percentage of materials uploaded by an individual user will be ephemeral or restricted to use by that individual. On the other hand, some materials will be thrown open by their owners to wider access. How will these items get cataloged? According to what standards? Who will provide guidance and training to contributors or catalogers? Will there be (an) organization(s) to "accession" some materials for permanent archiving and uniform access? Policy issues such as these are addressed in more detail in Section V.

Conclusions and Summary

The digital library projects we have surveyed in this document represent a number of different approaches to storing, managing, and delivering digital resources to their users. In concluding this section, let us take note of some important points:

1. The predominant model for the academic digital libraries we have surveyed is collection-focused.
2. The proposed mission for the UNC digital library is closer to that of the non-academic libraries we have examined, which are more resource-focused.
3. A resource-based focus allows for the creation of ad hoc collections or the retrieval of individual resources.
4. The most successful digital libraries employ rigorous standards in the creation of both their digital resources and metadata. This means those projects need not rely upon any particular media management solution and can migrate from one to another if necessary.¹⁸

While there are a number of digital library initiatives underway at UNC, none of them fulfills the function that we envision for this project: the management and delivery of digital resources to support the disparate educational and research missions of the university in a comprehensive fashion. Some of them fulfill this role for a specific discipline, others perform some of its potential functions, but most are area-specific collections of material.

Point #4 implies that storage and management should be thought of as independent of delivery in a logical sense, even if a single, end-to-end solution is employed. There are a number of solutions that might be employed to build a digital library to support teaching and research at UNC, but the fundamental standards and structure of the library should not depend on a particular one. In the next section, we will outline the committee's recommendations for the design of a UNC digital library and for standards and best practices for the creation of digital materials

¹⁸ A particularly good example of this is the Perseus Project, which, because of its concentration on standards-based resource creation, was able to migrate from a delivery system which relied on Apple's HyperCard on CD-ROM to a Web based delivery system without having to reinvent itself. See Gregory Crane, *The Perseus Project and Beyond*, D-Lib Magazine, Jan. 1998, <http://www.dlib.org/dlib/january98/01crane.html>.

IV A. Proposed Digital Library Specifications

In this section, we will outline what the committee thinks a UNC digital library should look like and how it should function. These specifications are not intended to be dependent on any particular solution or implementation.

Multimedia support

The design of a digital library for UNC must recognize the current needs of its potential users and be able to adjust to changes in those needs. In order for the digital library to be immediately useful we must identify what multimedia people are using right now on campus; but in order for the system to survive it must be capable of handling more digital media types, some of which may not even exist yet.

Any digital library solution, whether it is purchased from a vendor or developed in-house must be flexible enough to handle any number of multimedia types. The digital library should treat its multimedia content as objects, disconnecting the type of media from implementation of media. For example, we might recommend that users create TIFF image files or perhaps JPEG image files, but the system itself should recognize these files not by their implementation (TIFF, JPEG, etc.) but only by their typing: image. As a new image or audio type becomes more accepted on campus (as MP3 has recently become popular as a format for audio files) or even as a faculty member creates a new way of storing data, the digital library system must have the flexibility to bring that new media type on-line immediately.

Support for common tools

Users of a digital library system must be allowed to continue using as many of the tools with which they are familiar as possible. For example, a number of faculty members are currently using Blackboard's CourseInfo to create and publish course materials on the Web. Any campus digital library system should not require that such faculty members abandon their use of CourseInfo in order to be able to use the digital library. Instead the digital library should work in concert with such existing tools.

This policy should not only apply to Web-integration systems. For example, users who are comfortable with such products as Microsoft PowerPoint or Word should be able to utilize the resources of the digital library to populate their presentations and publications. Campus users must be able easily to integrate its new functionality and resources into products with which they are already familiar. We therefore recommend that any digital library system adopted by the university support legacy software and systems in order to take advantage of current user knowledge.

Integration with desktop environment

In a similar fashion, the digital library system should also integrate tightly with users' desktop systems, so that adding items to the library be as painless and intuitive as possible. Tools that require a high learning curve and offer a foreign, abstruse interface will not be readily adopted by the faculty.

It is our recommendation that the interface to the UNC digital library run on an established, universal platform. The only interface common to all computer systems across campus (and the world) is the World Wide Web. With a Web-based system, the digital library will offer immediate utility and can gain widespread acceptance on campus. There may in

addition be a need for integration with, e.g. the Windows desktop, but we see the Web interface as a baseline requirement.

Integration with campus systems

Any digital library system purchased or developed by UNC must support integration with pre-existing campus systems, such as the campus distributed file storage system (AFS). In order for the digital library to thrive, as much of the system as possible must be managed in existing and supported environments. It is equally important not to reinvent or reinvest in technology that is readily available on the UNC campus. For example, instead of managing userids and passwords within the digital library itself, there is already an authentication system in place that matches valid userids and passwords in the UNC email system. There is good reason to use this established “back office” system instead of attempting to create a new one, the most important being the active maintenance of user accounts within the current system. Integration with other systems, such as those run by AIS (Administrative Information Services) could allow the digital library to look up class rolls or perhaps charge for on-line course packs.

Tight integration also means that legacy systems can be supported and in some cases absorbed by the digital library—a primary goal of the original grant charter. A major problem that faced the granting committee was the disparity of campus systems originally funded under Chancellor Hooker’s 1996 and 1997 computer infrastructure grants. Many such systems had languished because no further funding was available after the original grant period. And many of these systems were incomplete in either form or function. One major purpose of a UNC digital library would be to unite and support such pre-existing systems, in order that the original grant money and the work it funded not go to waste. The committee does not currently have a recommendation on which systems in particular should be supported and/or absorbed, but the policy should be straightforward and should depend directly upon the availability of support for the project in question.

Resource discovery

Resource discovery is an important aspect of any digital library system. In order for a digital management and delivery system to be truly useful, there must be clear paths to the information users need. On the World Wide Web, there are two standard ways of accessing data: searching and browsing. Searching, whether by keyword or by more specific categories of information, is the direct pursuit of information using targeted words or phrases. Browsing by following links from one resource to another, on the other hand, may be purposeful or casual. For many users, browsing is a more intuitive and comfortable way to discover resources, particularly if they are searching for information about an area with which they are not familiar.

It is helpful for a user to know what a library contains before she browses or searches for information. A UNC digital library should include in its resource discovery mechanism a means of identifying the individual collections housed or indexed by the library. In short, a digital library should offer as many means of accessing its information as are thinkable, the most basic of these being both active and passive (browsing) search mechanisms. The library should also identify itself and its contents clearly and aid users in external resource discovery.

Data presentation

The digital library system should have the means both to return and format data in the style and/or format chosen by the user. The system should allow for arbitrary groupings of data,

as well as arbitrary ways of displaying data. For example, a professor using the system should be able to set up a series of resource collections for his current classes. These series might allow for two different views of the data, one for in-class presentations and one for the Web site where students can review the same information. These displays need not be limited to a Web interface; they might include, e.g., PowerPoint presentations or PDFs. It is therefore essential that the design of the system's data be separated from the presentation of that data. It is the committee's recommendation that a digital library support a multi-format, multi-display interface to the data, so that users can create multiple pathways to and views of the same data.

Security

Security is a central issue in digital resource management for materials that are not in the public domain. The system must be able to keep rights information about each resource it knows of and, when necessary, use that information to restrict access to resources. It is important to note that the issues of rights and restrictions are separate: rights information must be kept as part of the metadata attached to any given resource (see section IVB on metadata recommendations). Access to a digital object may be restricted in a number of ways either as an automatic extension of the owners' rights or as the result of policy decisions made by the UNC community. In other words: rights are legal statements intended to direct the behavior of potential users of an object, they cannot force those users to obey the law. Restrictions are rules governing how the management and delivery of certain objects may function, and they do enforce user behavior.

For example, a particular digital image may be copyrighted, and the copyright might state that users may only copy and distribute low-resolution versions of the image, if the copyright statement remains attached to the copies. There is nothing (other than the threat of legal action) to prevent a user from making any number of copies of that resource, and distributing them as he wishes. The same image may exist in both a high-quality, high-resolution version and a lower-quality version. It could be restricted so that users are allowed to access the low-resolution version at will, but are only allowed to know that the higher-resolution version exists (and whom to contact to acquire a copy) without being allowed to access it. The copyright is thus a further piece of information about the resource, while the restriction actually affects the functionality of the delivery system.

In order for some of these restrictions to function, the digital library must "know" with whom it is dealing. It should be able to tell, for example, whether or not a person is a member of the UNC community and adjust its behavior accordingly. Ideally, it should be able to tell whether, for example, a student should be allowed to access a particular resource because she is enrolled in a particular class. For this type of granularity to be feasible, the system would have to be tightly integrated with the existing computing infrastructure at UNC. It should be able to utilize existing authentication schemes (such as email username and password) and be able to retrieve data like class rolls from the AIS databases

Collaboration, user management, and workflow

An important and highly-requested feature of a UNC digital library is the ability to collaborate with colleagues on projects through the digital library interface. (For further information on the desired functionality of a UNC digital library, please see the results of the Fall 1999 survey, in Section II.) It was also made clear that this collaboration should be both internal and extramural. This collaboration, both inside and outside the university, would look very much like a Web-based discussion thread. Expert and student users alike could access any

given object in the system, adding their comments and amendments for public or private consumption. On the other hand, the digital library system should be flexible enough to allow for other collaborative efforts such as video conferencing and shared whiteboards.

Workflow represents the history of an object from its creation to its final form, along with all related comments, manipulations, and proxies. A workflow system allows for functions to be initiated by events in the life of an object. For example, an instructor might be notified when a student submits a paper. The student would in turn be notified when the instructor has finished annotating the paper, and so on. FACET (see section III, above) is an example of a system that manages workflow. We recommend that any digital library system support both collaboration between users and workflow.

Metadata

The digital resources stored on the World Wide Web can all be considered as information. A digital image of a Renaissance painting like the one to the right from the National Gallery of Art, therefore, is just another piece of information. The simple act of publishing that image to the Web, however, is often not sufficient to make it truly useful. More information may be necessary to help those who may wish to view the image or use it, such as its title, creator, and date of creation. Such information about a resource is called metadata. The uses of metadata are various and it is absolutely vital to the success of any library, whether physical or digital. Metadata may be used for descriptive purposes, to impart information about a resource. It can also play an important role in resource discovery if it stores subject information. In the example above, the picture's metadata might include keywords which would help someone searching for it to discover the resource. It might also contain a copyright statement and provide information about how the resource can be used by interested persons. Information about the ownership and location of the original painting could be included. A further use of metadata is for administrative purposes: it may be useful for the owners of the resource to keep information about how the digital resource was created. Was it made with a digital camera or scanned from a photograph or slide? What were the scanner settings? Was any post-processing done after the image was scanned? Who was responsible for its creation? They might go further and include information on when the cataloging record (the metadata itself) was created and by whom. The latter kinds of information help the owners of digital resources control the quality of those resources.

Example

Fra Angelico and Filippo Lippi

Fra Angelico:
Florentine, c. 1400-1455
Filippo Lippi:
Florentine, c. 1406-1469

The Adoration of the Magi, c. 1445
tempera on panel, diameter: 1.372 m
(54 in.)
Samuel H. Kress Collection
1952.2.2



Some uses for metadata:

1. Descriptive information.
2. Finding aids.
3. Rights and access information.
4. Information about how the resource was created.

It is clear from the discussion above that it may be desirable to keep all kinds of information about a given digital resource. For this information to be useful, it is best kept in a structured format, such as a database, so that it can quickly be retrieved when needed. It should also be clear that keeping such detailed information is not a simple task. This is why libraries devote so much time, attention, and expertise to cataloging. In general, the more information that can be captured about a resource, the more useful that resource will be. But, as we have seen, it is unreasonable to expect all potential users of a digital library to devote the time

necessary to create truly robust metadata. There must therefore be a compromise of some sort between completeness and convenience and the digital library's scheme for storing metadata must be able to handle both minimal and complete metadata. Our proposed schema uses XML to encode the metadata in a format that is based on the Dublin Core 1.1 standard. We believe this will allow the maximum flexibility in both storing and presenting the data.

Why XML?

XML (Extensible Markup Language) provides a very flexible mechanism for structuring metadata. XML is a standard for defining markup languages which has been recommended by the World Wide Web Consortium (W3C for short).¹⁹ The appearance of XML will be somewhat familiar to anyone who has seen or used HTML. The content of an XML document is contained within tags, or elements, which are marked off with angled brackets (<>). XML is designed to be easy to process using software and also to be understandable to human beings.

Some example XML

```
<?xml version="1.0" encoding="UTF-8"?>
<object objectid="1">
  <title titlequalifier="main">The Adoration of the Magi</title>
  <creator agenttype="person" agentrole="painter">Fra Angelico</creator>
  <creator agenttype="person" agentrole="painter">Filippo Lippi</creator>
  <description>tempera on panel, diameter: 1.372 m (54 in.)</description>
  <date datatype="Created" dateaccuracy="approximate">1445</date>
</object>
```

XML shares some of the advantages of other structured data formats, such as relational databases. An XML document can be checked for validity against a schema, thus ensuring that it contains all of the required elements and that those elements contain valid data. It is also capable of expressing complex relationships between data elements. XML's real advantage over other data storage formats is its portability. XML markup can be transformed into other types of markup, such as HTML, very easily. So an XML document can be delivered to a Web browser as HTML. The same document can be viewed in a wide variety of ways because of this transformative capability.²⁰

The advantages of XML:

1. Flexibility.
2. Consistency.
3. Portability.

Dublin Core

The Dublin Core is the result of an international effort to produce a standard set of elements for the description of electronic resources.

The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organizations.

The Dublin Core Workshop Series has gathered experts from the library world, the networking and digital library research communities, and a variety of content specialties in a series of invitational workshops. The building of an interdisciplinary, international consensus around a core element set is the central feature of the Dublin Core. The progress represents the

¹⁹ See <http://www.w3.org/TR/1998/REC-xml-19980210> for the specification for XML 1.0. An annotated version of the specification is available at <http://www.xml.com/axml/axml.html>.

²⁰ XML transformation is accomplished via XSL, a transformation markup language which is itself written in XML. See <http://www.w3.org/Style/XSL/> for detailed information.

emergent wisdom and collective experience of many stakeholders in the resource description arena.²¹

The standard is designed to be both simple and extensible. DC 1.1 consists of fifteen optional, repeatable elements, several of which may be qualified in various ways (for example, a description may be identified as an abstract or a table of contents).²² The proposed scheme is based on this standard.

The Dublin Core Element Set

Element	Qualifier(s) ²³
Title	"Alternative"
Creator	
Subject	Encoding Scheme ²⁴
Description	Description Type
Publisher	
Contributor	
Date	Date Type
Type	Encoding Scheme
Format	Extent, Medium, Encoding Scheme
Identifier	Encoding Scheme
Source	
Language	Encoding Scheme
Relation	Relation Type, Encoding Scheme
Coverage	Place, Time, Encoding Scheme
Rights	

The Proposed UNC Schema

Element	Qualifier(s)
Title	titlequalifier
Creator	agenttype, agentrole
Subject	vocabulary
Description	externaldesc descriptionqualifier
Publisher	agenttype, agentrole
Contributor	agenttype, agentrole
Date	datatype, dateaccuracy
Type	
Format	formatscheme, formatresolution, filesize
Identifier	identifierscheme, identifierid
Source	
Language	

²¹ From the homepage of the Dublin Core Metadata Initiative, <http://purl.org/DC/>.

²² See <http://purl.org/DC/documents/rec-dces-19990702.htm> for the Dublin Core Element Set.

²³ The Dublin Core Usage subcommittee has recently approved a set of qualifiers to the basic element set. These qualifiers are meant to aid in the interpretation of element contents and / or make those contents more specific. The official recommendation is not available at this time, but the announcement may be found at <http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html>.

²⁴ From the announcement cited above: "These qualifiers are pointers to schemes that aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations. A value expressed using an Encoding Scheme will thus be a token selected from a vocabulary (e.g., a term from a classification system or set of subject headings) or a string formatted in accordance with a notation (e.g., "2000-01-01" as the standard expression of a date)."

Relation	relationtype
Coverage	coveragetype, coveragescheme, coveredatum
Rights	
Meta	

The proposed schema shares a great deal in common with the Dublin Core, with the addition of some qualifiers, and two additional elements. We also propose extending the Coverage and Rights elements with the addition of subelements. In our scheme, the Coverage element might contain a Coordinate element. This will allow us to include detailed geographical or temporal information. The Rights element can contain three additional subelements: Rightsstatement, Rightscontact, and Rightsrealm. The first of these is a container for, e.g. a statement of copyright on the resource being described. Rightscontact might provide the email address or telephone number of a person who can be contacted for permission to, e.g. make a copy of the resource. Rightsrealm is intended to extend the Rights element to allow for authentication and access management for resources. It contains the ID number of the persons or groups who are allowed to access the resource and a list of access controls.

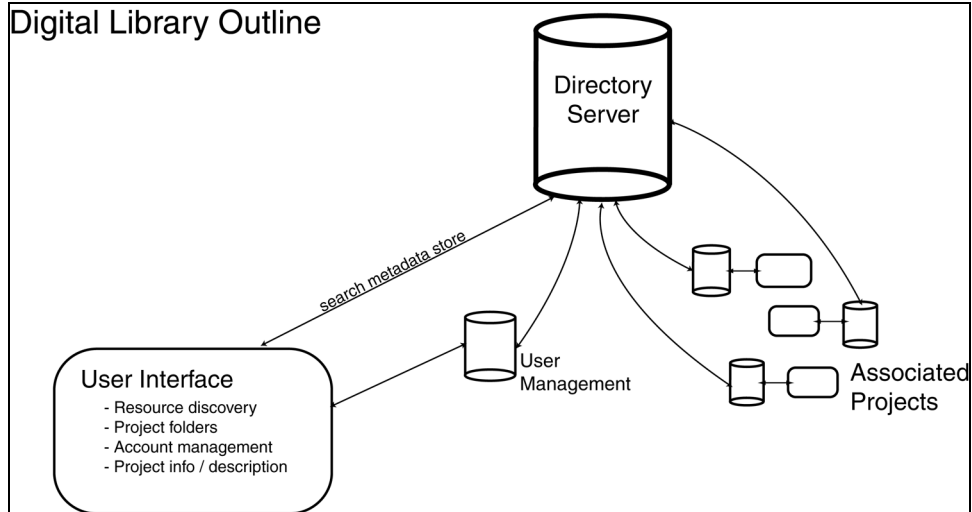
The new element in our schema is Meta. This element is designed to be a container for information that cannot be mapped into our schema. Individual disciplines may have special needs for types of information that are different from or more detailed than what the DC standard supports. The Meta element is intended to allow this type of information to be stored and accessed as required.

Further information (including the Document Type Definition we have developed and some sample files) is available in Appendix A of this report.

An outline of the proposed system

The UNC Digital Library Project and the Triangle Digital Libraries Group have produced an outline of how a digital library to support teaching and research might look. We envision at least two systems, one to manage resource discovery, and the other resource delivery and manipulation. The first system, a directory server, would contain descriptive metadata about all of the digital resources associated with the digital library. It would allow for a variety of different types of search and would support access restrictions of various sorts (for example, allowing users to discover the existence of an object, but not providing them with it, if they lacked the necessary permissions). The second system would allow users to manipulate resources discovered via the directory and perhaps also external resources. It would permit users to assemble collections of digital material, to annotate those materials and collections, and to secure or permit access to those collections by other users. It would thus provide mechanisms both for delivering digital content in the context of a class and for sharing material with colleagues. Further components might include centralized storage facilities for digital material and digital image acquisition stations.

The diagram below illustrates how the components of this system might interact.



Summary

The committee has several general recommendations for the UNC digital library in the areas of resource management and delivery. We believe that if the digital library is to proceed to the implementation stage it should:

1. Utilize and be able to integrate with existing UNC systems.
2. Support tools and platforms commonly used by the UNC community.
3. Provide multiple paths for users to discover resources.
4. Support multiple presentation formats for data, from slide shows to “digital course packs.”
5. Support collaboration between users within the UNC domain, and, if possible, outside the university as well.
6. Implement robust security, where necessary, to protect copyrighted materials and intellectual property.
7. Follow open, published standards in its design.

In addition to these considerations, there are a number of policy decisions that must be made in order for a centralized resource management and delivery system to function.

1. There should be clear guidelines for dealing with copyrighted materials in a digital library setting, so that the system can implement the appropriate access restrictions.
2. There must be decisions made about how the various digital library systems are to be managed and supported.
3. The issue of storage is particularly important. There will need to be policies in place for handling very large quantities of digital material in a secure and robust environment.

IV B. Standards

The following recommendations are directed towards members of the UNC community who are producing, or plan to produce digital resources for use in teaching and research. For those who are considering the production of digital resources, it may be helpful to answer the following questions before proceeding:

1. What is the goal of your project? Are your requirements for the digital resources you will be producing strictly short-term, or long-term? The answer to this question will help determine whether you should produce archival-quality master files. If you wish the resources you are producing to have more than a short lifetime, then you should certainly produce archival masters, which can be used to generate surrogates for delivery in a variety of situations.
2. Are there copyright restrictions on the original objects you will be digitizing? Will these restrictions affect your use of the digitized resource? If the original is copyrighted, it is probably best to avoid making archival masters, unless there is some overriding reason to do so, and even then it will be necessary to restrict access to the digital copies.
3. Is the quality of the originals such that it is worth keeping high-quality reproductions? Pick out a set of good originals that will satisfy the purposes of your project. Good clean originals give you higher quality digital copies and are more efficient in the use of storage space.
4. Is there a system in place for keeping information about the digital resources? The value of a resource may decrease substantially if that resource is separated from information about it. How will you capture information about each resource, and how will you store it?

In general, we would recommend the creation of archival masters for most digital resources. These are files that may be too large to be conveniently delivered to users now, or may be in formats that don't lend themselves to direct distribution, but which may be used to create surrogate copies for use in many contexts. For resources whose longevity is a concern, archival masters are an absolute necessity. Therefore our recommendations will, in many cases, be divided between advice for archival and "ephemeral" versions of your digital media.

Images

There are already recommended standards for digital images published on the Web by the Library of Congress (for the American Memory Project)²⁵ and the California Digital Library.²⁶ Our recommendations are derived from these. There are a number of file formats in common use for storing images, such as TIFF (Tagged Image File Format), JPEG (Joint Photographic Experts Group), JFIF (JPEG File Interchange Format—a commonly used implementation of the JPEG standard), GIF (Graphics Interchange Format), and PNG (Portable Network Graphic). TIFF, GIF, and PNG use lossless compression algorithms, which means that no information is discarded when a file is compressed. JPEG's compression is lossy, but results in much smaller files, which is why it has become the standard for continuous-tone images on the Web. TIFF, JPEG, and PNG support 24 bit (or higher), color, which means they are able to represent continuous-tone images, such as color photographs without altering the original's color scheme.

²⁵ <http://memory.loc.gov/ammem/formats.html>.

²⁶ <http://www.ucop.edu/irc/cdl/tasw/Current/Imaging.Stds-090199/Imaging.Stds-090199.pdf>.

GIF does only 8 bit color, and is thus not suitable as a storage or delivery format for high color images.²⁷

Image resolution is typically defined in dots or pixels per inch (dpi or ppi). It should be noted, however, that this type of measurement is more useful for determining the quality of an image when it is printed than when it is viewed on a monitor. For purposes of on-screen display, what matters is the number of pixels along each side of the image—this will determine how large the image appears on a computer screen.

We recommend that persons wishing to create digital images take the ATN Training Center classes dealing with Adobe PhotoShop. Information on these classes and links to the current class schedule may be found at <http://help.unc.edu/training/descript/image.html>.²⁸ There are a number of Web pages devoted to scanning. <http://www.scantips.com/> has the fullest and most up-to-date explanations we have found.

Archival

Archival master images should be stored in a format that does not discard any information during file compression and which is capable of storing the image in 24-bit color. The most commonly used such format is TIFF. If the preservation of the color of the original is important, images should be scanned with a color bar or corrected against the original. Likewise, a scale bar should be included if the size of the original object is an issue.

minimal standards:

35mm film

Tonal depth: 24 bit
Resolution: 2800 ppi, i.e. 4200 pixels on the longest side for 35mm film.
Format: TIFF or PNG
Compression: Uncompressed, LZW, or PNG compression

medium to large color photograph or transparency

Tonal depth: 24 bit
Resolution: 3000-6000 pixels on long side.
Format: TIFF or PNG
Compression: Uncompressed, LZW, or PNG compression

bitonal text document

Tonal depth: 8 bit
Resolution: 600ppi, e.g. 4800 x 6600 pixels per side for a letter-sized page.
Format: TIFF or PNG
Compression: Uncompressed, LZW, CCITT Fax 4, or PNG compression

²⁷ For detailed information on these and other image file formats, see the Graphics File Format page at <http://www.dcs.ed.ac.uk/home/mxr/gfx/>.

²⁸ The documentation for these classes may be found at <http://help.unc.edu/cgi-bin/getindex?doctype=Desktop&platform=Graphics%20Applications&type=html>. We particularly recommend the appendix to the PhotoShop: Image Acquisition handout, at <http://help.unc.edu/cgi-bin/getdocs?docnumber=dap02&doctype=Desktop&platform=Graphics%20Applications&product=Photoshop#Heading10>.

Ephemeral

Non-archival images may be stored at a number of different sizes, resolutions, and formats, depending on the requirements for that file. For example, there may be a thumbnail version and a larger version for viewing details.

high-color images:

Large

Tonal depth: 24 bit (color)
Resolution: 600–1800 pixels on the long side of the image.
Format: JFIF
Compression: JPEG²⁹

Thumbnail

Tonal depth: 24 bit (color)
Resolution: 150 pixels on the long side of the image.
Format: JFIF
Compression: JPEG

low-color images

Large

Tonal depth: 8 bit
Resolution: 600-1800 pixels on the long side of the image
Format: GIF or JFIF

Thumbnail

Tonal depth: 8 bit
Resolution: 150 pixels on the long side of the image
Format: GIF or JFIF

Text

Archival

For archival versions of texts, we recommend avoiding proprietary formats such as Word, WordPerfect, and PDF. We also recommend avoiding formats that do not properly encode the structure of a document, but instead emphasize style, such as RTF and HTML. There are well-defined standards for marking up texts in SGML and XML for a number of disciplines, and these are the only truly viable ways of creating archival masters for texts. Once a document has been marked up, it can be used to generate surrogates in a variety of formats (such as HTML, PDF, etc.) for delivery to users.

archival texts:

Format: SGML or XML
DTD: If there is a standard DTD (Document Type Definition) that pertains to the particular type of document, that should be used (e.g. TEI,³⁰ MathML,³¹ etc.).

²⁹ A number of image editing programs, such as Adobe PhotoShop, allow users to choose the level of compression used when saving an image. PhotoShop compression ranges from 1 to 10, 1 being most lossy and 10 being virtually lossless. We recommend using compression levels above 6 for most images.

³⁰ <http://www.tei-c.org/>.

³¹ <http://www.w3.org/TR/REC-MathML/>.

Ephemeral

We expect ephemeral texts to be stored in a variety of formats, but recommend that, as far as is possible, metadata about these documents be stored within them. Most formats (such as Word, PDF, etc.) allow users to note who created documents in the properties of their files. Adobe's PDF (Portable Document Format) format, although proprietary, is very convenient. Its specifications have been published, and the software for viewing PDF documents is free. PDF allows users to view and print documents exactly as they were formatted by their creators and it has become a *de facto* standard on the Web. For HTML, we recommend that data such as creator and a description of the document's content be stored in META tags in the document header.³² HTML should be compliant with one of the W3C standards.³³ This can be accomplished automatically with tools such as Tidy³⁴ or HTML-Kit,³⁵ both of which are free and will convert HTML with errors such as unclosed tags into valid HTML or XHTML.

ephemeral texts:

Formats: PDF or HTML

Audio

For both audio and video the standards are constantly evolving. We recommend a conservative approach to creating archival-quality media for both. In the case of audio, the quality of the digital master will very depending on the source. You will get higher quality sound from a CD than you will from monophonic tape. As with images, we recommend that you create uncompressed or losslessly compressed master files, from which other formats, such as MP3 or Real Audio can be derived for delivery to your users.

CD-quality audio

Sampling rate: 44.1 kHz
Resolution: 16 bit
Channels: 2 (for stereo sound)
Encoding: Pulse Code Modulation
Formats: WAVE, AIFF, or AU

Video

Since digital video is such a rapidly developing field, and since implementations of digital video depend to a great extent on the infrastructure and support available, we recommend that you contact the Center for Instructional Technology's Video Services group (<http://www.unc.edu/cit/vidserv/video-services.html>) for advice on creating digital video. Another important group dealing with digital video at UNC is the Center for Advanced Video Network Engineering and Research (<http://www.unc.edu/cavner>).

³² For a recommendation on how to use Dublin Core in HTML, see <http://www.ietf.org/rfc/rfc2731.txt>.

³³ See <http://www.w3.org/TR/html4/> for the HTML 4.01 specification and / or <http://www.w3.org/TR/xhtml1/> for the XHTML 1.0 specification.

³⁴ <http://www.w3.org/People/Raggett/tidy/>.

³⁵ <http://www.chami.com/html-kit/>.

Metadata

The recommendations for this section apply to all of the others, since you should be keeping information about any digital resources you create, no matter what the form of those resources. Each discipline that may wish to create digital resources for education and research will potentially have different metadata requirements. We recommend that any metadata scheme you employ be consistent and well thought out. There are a number of published metadata standards, some of them discipline specific, so we advise that you do some investigation of applicable materials before defining your own scheme. The following are some starting points:

1. Dublin Core: <http://purl.org/DC>
2. RDF (Resource Description Framework): <http://www.w3.org/RDF/>
3. IMS (Instructional Management Systems): <http://www.imsproject.org>
4. VRA (Visual Resources Association) Core:
<http://www.gsd.harvard.edu/~staffaw3/vra/coreinfopage.htm>

Some of these specify methods for packaging metadata, some do not. Some provide tools for working with metadata. All of them are the work of professional librarians, museum curators, educators, and computer professionals.

Archival

Metadata for archival resources must necessarily be more extensive than that for ephemeral objects. There are three categories of metadata that should be considered: descriptive, structural, and administrative. Some of these have been discussed above in the digital library specifications. One primary use of descriptive metadata is to aid in resource discovery. It may, however, also contain information that does not serve this function, but simply provides useful information about a digital object or its original.³⁶ Structural metadata, as its name implies, describes the structure of the resource. If it is a digitized book, for example, consisting of many images of the pages, the structural metadata might describe the number and character of those pages. Administrative metadata is particularly important when it comes to archival-quality resources. It may, for example, be very useful to know information about the lifecycle of a digital resource. How was it digitized and by whom? Was any post-processing done after it was digitized (such as color correction)? Who was responsible for creating the descriptive metadata? Some of this metadata (particularly administrative) may be generated automatically, during the process of resource creation. We strongly recommend that you seek the advice of a professional librarian before implementing any metadata scheme for archival material.

Ephemeral

Metadata is still important for ephemeral material, but there is likely to be less need for, e.g. administrative information. The amount of information you keep on any given digital resource will vary depending on how that resource is to be used. We suggest at a minimum that you keep information like title, creator, and creation date. The more such information you keep, the more useful the resource will be to others, and the easier it will be for others to find the resource.

³⁶ For example, information on the dimensions of an original will not help in a keyword search, but may still be extremely useful for someone looking at the resource.

V. Implementation

The following matrix compares the relative strengths and weaknesses of the two candidate commercial software products (Informix Internet Foundation 2000 + Media 360 vs. Oracle 8i + WebDB + *interMedia*) with respect to 14 major capability areas identified by the committee. At the same time, the matrix indicates the relative amount of configuration, customization or extension work that will be required to achieve the necessary functions and legacy-system interoperability for each capability area. Both products provide robust customization and configuration options using standard programming languages and interfaces. Nonetheless, deployment of either product will require “extension” work similar in nature and scope to the tasks a large business could expect to undertake in order to implement and accommodate its unique management processes, business logic, and corporate structures. Selection alone of either commercial product is insufficient to effectively implement a viable digital library system that will be maintainable and responsive to the needs of the UNC-CH teaching and research mission; therefore, assessment of the product trade-offs must be conducted in the context of the likely effort necessary to implement these extensions.

If UNC, Duke and NCSU continue to participate in and devote resources to the Triangle Digital Library Group (TDLG), that initiative can be expected to produce a significant portion of the logic, standards and resources necessary to support the interinstitutional and distance-learning functions proposed for the digital library system. For this reason, the “extensions” portion of the matrix has been subdivided to indicate the relative amounts of effort that can be expected to fall to the responsibility of UNC-CH alone as opposed to the effort a UNC/Duke/NCSU-supported TDLG effort can be expected to handle. Participation in, and integration with, the TDLG effort is likely to provide further synergies at the campus level (e.g., shared code, cooperative problem-solving and testing), promote the use of widely-accepted standards, and reduce implementation risk and timeline for mutually useful components.

For each capability area, colored bars are used to indicate the effort necessary to implement that capability. The matrix does not attempt to indicate the total amount of effort required from capability area to capability area (i.e., in terms of person-hours or dollars); rather, each bar is equivalent to 100% of the effort required for a given capability. Each “effort bar” spans 1-3 columns, corresponding respectively to the “native functionality” of each commercial product, the extension / configuration required at the UNC level alone, or the extension / configuration / implementation that is expected to spring from a UNC/Duke/NCSU-supported TDLG effort. If the choice of commercial product has no impact on where implementation effort must be placed, a single purple bar is used. If the distribution or level of implementation effort differs depending on the choice of commercial product, two colored bars are juxtaposed (blue indicating Informix’s package, red indicating Oracle’s). For example:

Capability, feature or requirement		Native functionality	UNC extensions	UNC + TDLG extensions
2.	Brews coffee			
3.	Fries potatoes			
4.	Calls fire department automatically			


















Capability, feature or requirement	Native functionality (in off-the-shelf candidate commercial products)	UNC extensions (support intramural needs; implement UNC logic & processes)	UNC + TDLG extensions (support inter-institutional / distance-learning needs)
1. Implements descriptive cataloging (metadata) scheme that complies with standards and facilitates the digital library's mission			
5. Absorption and interaction with legacy media systems			
6. Presentation (classroom & alternate learning environment)			
7. Publish to (course) Web site (including existing infrastructure)			
8. Support for collaborative sharing, submission and review of media and comments in an academic context (colleagues, students) <ul style="list-style-type: none"> • intramural • interinstitutional / distance learning 			
9. Multimedia management and indexing "inside" the system <ul style="list-style-type: none"> • MS Office file formats / Acrobat PDF • Image / Audio / Video • Web / HTML / XML • Geodata 			
10. Processing platform and server hardware			
11. Data storage (space for digital media and descriptive metadata)			

Informix video support is more feature-rich

Oracle XML support is more robust

Both products require add-on Informix add-on works with ESRI (already UNC site-licensed)

*Oracle: assume use of existing HW
Informix: can't coexist with Oracle-retasking or*

Capability, feature or requirement	Native functionality (in off-the-shelf candidate commercial products)	UNC extensions (support intramural needs; implement UNC logic & processes)	UNC + TDLG extensions (support inter-institutional / distance-learning needs)
12. Cross-browser & cross-OS client compatibility			
13. Security / access / rights management in "webbed" environment <ul style="list-style-type: none"> • intramural • interinstitutional / distance learning 	   	  	
14. Tight integration with workstation desktop and 3 rd -party tools			
15. Media workflow and lifecycle management			
16. Back-office integration (class rolls, userids, library, web servers) <ul style="list-style-type: none"> • intramural • interinstitutional / distance learning 		 	
17. Interaction with 3 rd -party servers (e.g., streaming media)			

Oracle provides more options, configuration still required

Informix support for Netscape still in development

Informix provides more features and options in both areas

Discussion of capability implementation effort

1. Implements descriptive cataloging (metadata) scheme that complies with standards and facilitates the digital library's mission

The implementation of a standard, descriptive cataloging scheme that will facilitate the discovery and use of digital resources has been identified as a top priority for any UNC digital library system. The topic comes up again and again: in reviewing international standards documents, talking to vendors, interviewing faculty and library staff, analyzing existing projects, during TDLG meetings, and in discussion with FITAC members. In particular, FITAC has stressed the importance of a metadata scheme and implementation that directly supports the instructional and research missions of the university while maximizing opportunities for interaction with legacy and interinstitutional systems as well as reuse of descriptive data created by existing digital media projects across campus. In cooperation with the TDLG, the committee and staff of the UNC digital library project have developed a candidate metadata scheme that would serve the needs of users across campus disciplines as well as a triangle-wide digital resource directory system.³⁷ Both vendor packages provide the architecture and functionality necessary to implement the UNC/TDLG metadata scheme.

- Informix provides more options for media lifecycle management and a more fully-developed work-flow environment than Oracle
- Significant UNC and TDLG effort will be required to develop, implement and test controlled vocabularies and discipline-specific requirements
- There must be long-term institutional support, broad adoption and long-term viability at risk

The area of descriptive metadata will also require some policy decisions. The creation of such metadata will, in the first instance, be the responsibility of the creator of a given digital resource. It will be difficult, however, to maintain any degree of consistency in the metadata with such a heterogeneous group of catalogers. Some users may be reluctant to supply complete information, because of the time involved. Some of the data entry work may be done by users (such as teaching assistants) who are not necessarily experts in the subject matter. These cataloging issues can be at least partially resolved by the user interface and by the development of discipline specific vocabularies for cataloging digital media. But the following issues must also be considered:

- What role (if any) should professional librarians play in the creation of metadata for the digital library?³⁸
- What is the minimum amount of information that must be supplied for a digital resource?
- How should the development of controlled vocabularies and discipline-specific metadata schemes be handled?
- Should there be policies and methods instituted for “vetting” descriptive metadata, such as having an expert in the subject area review catalog records?

³⁷ See above, section IVA, Metadata.

³⁸ Should some sort of collaboration with the library be instituted, whereby content creators supply some metadata while professional catalogers supply subject headings? See James Weinheimer's "Proposal for a Metadata System," <http://www.princeton.edu/~jamesw/mdata/metadataprop.html>, which suggests how this type of collaboration might work. Such a scheme would require additional funding for library personnel to handle the work.

2. Absorption and interaction with legacy media systems

It was interest in eliminating duplication of effort, guaranteeing the long-term viability of prior digital media investments and facilitating future use of digital multimedia that led to the funding and staffing of the UNC digital library project in the first place.

Success in merging and/or interoperating with a still-growing number of legacy digital media systems depends on two factors: scaleable core support for a range of media types, including those broadly used and licensed across the campus, as well as the implementation of a standard, flexible metadata scheme (see no. 1, above). These concerns are shared by digital library researchers and implementers on all three TDLG campuses, and significant energy has already been invested in addressing them. All participants agree that any approach to dealing with legacy systems must encompass both the ability to import and rehost the entire content of a moribund system as well as establishing long-term interoperability linkages with systems that continue to be managed, maintained and expanded by individual units, departments, projects and consortia.

- Both Oracle and Informix provide excellent support for required media types
- TDLG will produce standards and mechanisms for legacy interoperability
- Resolution of individual legacy situations at UNC will require significant effort

3 & 4. Presentation (classroom and alternate learning/research environment) and Publish to course Web site (including existing infrastructure)

Presentation and Web publishing are both aspects of the same issue: content delivery. Part and parcel with staff and faculty concerns over well-designed metadata is the imperative of providing quick, reliable and easy-to-use mechanisms for finding, selecting, organizing and annotating digital materials from disparate sources and then moving the aggregate result into the learning or research environment, whether that be classroom, library carrel, office, dorm room, coffee house, field location or distance learning situation. Classroom presentation, in particular, requires a solid infrastructure independent of the vendor solution employed.

- Both Oracle and Informix provide multiple, robust mechanisms for content delivery; however, these focus primarily on e-business, corporate communications and advertising/news media publication
- Specific functionality for academic delivery requirements must be added through UNC-local customization and integration of commercial product with existing delivery mechanisms (e.g., CourseInfo and course web pages)
- End-to-end academic delivery functions are essential for broad adoption and long-term viability

5. Support for collaborative sharing, submission and review of media and comments in an academic context (colleagues, students)

Collaboration and digital interaction are cornerstones of the most exciting of the emerging digital library technologies (e.g., Berkeley DLP). Flexible means of providing variable levels of authorial and review control over digital resources to arbitrarily constituted user groups (e.g., classes, research teams, interinstitutional committees) will help revolutionize many aspects of professional life in the academy.

- Informix's greater native support for media lifecycle and work flow management would facilitate implementation of these features more easily

- Support for inter-institutional collaboration at the level of individual objects depends on development of mechanisms for interinstitutional user authentication and resource rights management, issues currently under discussion in the context of TDLG.

6. Multimedia management and indexing “inside” the system

Support for many different types of digital media is another essential element of digital library technology. Faculty and staff using the digital library should be able to continue to work with the applications and file types to which they are accustomed.

- The advantage of Object Relational Database Management Systems (ORDBMS) such as the Informix and Oracle products is their ability to treat such diverse formats as PDF, Word, XML, and JFIF as native datatypes—so that they are able to perform functions like full text searches on Word documents, or find images that are similar to one another.
- Both Oracle and Informix have the advantage in different areas. Informix has better support for working with digital video and its GIS DataBlade works with ESRI software, which is already licensed by UNC. Oracle’s support for XML is more thorough.

7. Processing platform and server hardware

The issue of allocating hardware space to the digital library is mainly a matter of policy, but two points should be noted:

- Since UNC already has a site license for Oracle, there is a platform in existence for a prototype digital library to run on.
- A decision to use Informix must be based on a commitment to acquire or retask a server for that software.

8. Data storage (space for digital media and descriptive metadata)

Data storage is another area that will require policy decisions. Some types and formats of digital media require very large amounts of storage space. A digital library using Informix or Oracle might deal with data storage in two ways, internally or externally. If the storage is internal, the digital resources are actually stored inside the database, which has several advantages, but also means the DBMS and the platform it resides on must be able to support the volume of data required. With external, or distributed, storage, the database stores only pointers to the actual files and metadata about them. In the latter case, the files may reside anywhere on the network. In at least some cases, files will have to be stored externally, either because they belong to existing projects or because their owners insist on storing them on hardware that they control. The following decisions about data storage must be made before the digital library can be implemented:

- Will any digital media be stored inside the database?
- How will expanding needs for storage space be handled? Will individuals or departments be asked to buy extra storage space as needed?
- How much space will be allocated to users and where will external space reside?

9. Cross-browser & cross-OS client compatibility

With the Carolina Computing Initiative, the desktop computing environment at UNC has become more homogenous, but it is still quite diverse. The committee has recommended that any digital library support a Web interface which functions in any recent browser. If such

support is implemented, differences in operating system should have no effect on the user's ability to exploit the capabilities of the digital library.

10. Security / access / rights management in "webbed" environment

The implementation of access management in the digital library is crucial to its success. There must be support for controlling access to materials that are copyrighted or sensitive. No vendor digital library solution is capable of resolving this issue, because it depends upon a number of factors outside the database. The chosen product should be able to be integrated into a campus-wide authentication scheme and should support the following:

- Single authentication (where possible) for access to all permitted content.
- Different levels of access permissions, such as "discover," "view," "modify," "remove."

11. Tight integration with workstation desktop and 3rd party tools

As we noted above (#6), it is important that faculty using the digital library be able to continue to use the tools they are accustomed to.

- Informix's use of 3rd party DataBlades often provides better support for interaction with other software.

12. Media workflow and lifecycle management

The ability to manage workflow is a particularly desirable function for a digital library. This might include, for example, alerting a faculty member when her colleague has made a change to a digital resource that is an element in a collaborative project, or alerting a graduate student when his advisor has finished annotating a dissertation chapter.

- Informix's Media360 has built-in workflow management tools.

13. Back-office integration

It will be desirable for the digital library to perform functions such as authenticating members of a particular class for access to that class's materials, while preventing others from seeing them. This type of functionality depends on the library's ability to be integrated with existing UNC systems and must therefore be developed in-house.

14. Interaction with 3rd party servers

Some media types, particularly video, may have to be served from a machine dedicated to that purpose. In this kind of situation, the digital library must be able to interact with that server (e.g. by passing authentication information). Both vendor products allow for this kind of interaction.

APPENDIX A: Committees and Staff

UNC Digital Library Project

Co-Leaders:

Steve Weiss (Computer Science), Bob Henshaw (ATN / CIT)

Committee:

Gary Marchionini (SILS), Toby Considine (Facilities Services), Carl Ernst (Religious Studies),
Russ Van Wyk (Cary Academy), Celine Noel (Academic Affairs Library)
David Warshauer (Medical School), Julia Shaw-Kokot (Health Sciences Library)

Staff

Hugh Cayless (Project Manager), Noel Fiser (Graduate Assistant, Classics)
Tom Elliott (Graduate Assistant, History)

Triangle Digital Libraries Group

Duke

Jim Coble, John Little, Paolo Mangiafico, Randy Riddle

NCSU

Caroline Beebe, Bill Colman, Keith Morgan, Andrew Pace, Shirley Rodgers, Rob Rucker,
Deborah Westmoreland

UNC

Hugh A. Cayless, Phyllis Daugherty, Tom Elliott, Noel Fiser, Bob Henshaw, Andy Ingham, Gary
Marchionini, Celine Noel, Julia Shaw-Kokot, Tim Shearer, Sandra Shirley

Triangle Research Libraries Network (TRLN)

Mona Coutts

APPENDIX B: Digital Library Glossary

This glossary was created in order to educate the digital library committee on some of the key terminology related to digital libraries.

API

Application Programming Interface. An API is the set of functions a piece of software, an operating system, or a piece of hardware provides to allow software to talk to it.

CGI

Common Gateway Interface. A process which runs on a server and allows you to program a dynamic web interface. CGI is one type of middleware. Standard HTML pages are static--they are documents created by a web developer and the same documents are read by users. CGI makes it possible for web pages to respond to user behavior. If you've used a search engine, you've probably filled in a form and then posted it by hitting a button. A CGI script then takes the information from your form and sends it to the site's database as a query. The database responds with whatever information the query returns, and the CGI script then writes out an HTML page containing those results.

client/server architecture

A network architecture in which every machine is either a client or a server. Typically, servers perform tasks that involve managing network resources. For example, a file server provides storage space and access to that space over the network. A Web server provides access to files over a network too, and may run middleware programs that provide server-side processing.

client-side

In client-side processing, the data processing takes place on the client itself. For example, the results of a query might be delivered entire to the client, which would itself format those results for display in a Web browser.

CSS

Cascading Style Sheet. A set of instructions that can be used to format HTML or XML files. These instructions specify how text and other objects within markup tags are to be displayed. For example, you can use a stylesheet to modify the way text within paragraph (<P></P>) tags is displayed, so that the first line is indented and the text is displayed in a particular font.

database

A database is any organized collection of data (e.g. a telephone directory). Often, but by no means always, databases are created and managed using software specially designed for the purpose (DBMS's).

DBMS

Database Management System. A software package that allows users to work with a database. DBMS's typically provide functionality like creating and modifying tables, adding and deleting fields and records, and retrieving data with queries.

DOM

Document Object Model. An API for manipulating XML documents. The DOM functions by creating an in-memory tree from a given XML document. Because it relies on loading the whole document into memory, it may not be efficient for very large documents (See SAX).

DTD

Document Type Definition. A DTD prescribes the tags available for marking up a particular class of SGML documents, so HTML is HTML because it has a DTD which prescribes the tags which can be used to create HTML markup. See <http://www.w3.org/TR/REC-html40/sgml/dtd.html> for a DTD for HTML 4.0.

EAD

Encoded Archival Description. An implementation of SGML designed for use by museums, archives, libraries, etc. to facilitate the electronic cataloging of collections.

fat client

In a client/server architecture, a fat client is one that performs the bulk of the data processing itself.

GIF

Graphics Interchange Format. GIF (pronounced jiff, or gif) is an image format which supports only 8-bit colors (256 colors), so it is most useful for images without a high degree of color gradation. GIF89a (the latest version of the format) allows for transparency and animated images.

HTML

Hypertext Markup Language. HTML is the authoring language used to create documents on the World Wide Web. HTML is similar to SGML, although it is not a strict subset.

JDBC

Java Database Connectivity. An API that allows programs written in Java to communicate with databases.

JPEG

Joint Photographic Experts Group. JPEG is a standard for compressing digital images. The name (pronounced jay-peg) comes from the committee who wrote the standard. JPEG is a “lossy” format, which means that it uses a compression algorithm which discards some of the original information. How much information is discarded depends on the parameters chosen when the file is saved. JPEG is one of the standard formats used on the Web. It supports 24-bit color (over 16 million colors), so can be used for pictures with lots of color gradation. GIF, on the other hand allows for only 256 colors.

metadata

Metadata is “Data about data.” Its purpose is facilitating the task of locating relevant documents (be they books, digital files, slides, etc.). See Introduction to Metadata; Pathways to Digital Information (<http://www.getty.edu/gri/standard/intrometadata/index.html>), from the Getty Information Institute.

middleware

Middleware is software which enables two applications to communicate with one another. On the Web, middleware typically performs functions like allowing information to be passed back and forth between a Web browser and a database. For some examples of Web middleware, see <http://idm.internet.com/tools/webdev.shtml>.

ODBC

Open Database Connectivity. An API for communicating with databases. ODBC allows an external program to enter, retrieve, or manipulate data in a database. For example, it might allow a CGI script to retrieve data from a database for display on a Web page.

OLAP

On-Line Analytical Processing. A category of software tools that provides analysis of data stored in a database. For example, OLAP can analyze trends and summarize large amounts of data. OLAP may be contrasted with OLTP, which deals with one record at a time and is concerned with processes. OLAP is concerned with many records at a time, and is subject-oriented (e.g. it is concerned with questions like “How is the database being used?”, or “What objects are most/least popular?”)

OLTP

On-Line Transaction Processing. A type of processing in which the system responds immediately to user requests. Each request is considered to be a transaction. OLTP is the opposite of batch processing, in which a number of requests are cached and then processed all at once.

OODBMS

Object-Oriented Database Management System. Also abbreviated ODBMS. A system which combines the characteristics of a DBMS with those of an object-oriented programming language. A DBMS stores data as objects rather than as fields and records in tables. Object-oriented databases have some advantages over traditional RDBMS's including the ability to create new and/or complex data types, and the ability to model complex relationships between data. This is a relatively new technology, however, and it lacks some of the features of the better-established RDBMS model, including a well-developed query language like SQL.

Open Source

The Open Source Initiative (<http://www.opensource.org/>) certifies software as Open Source if it meets the consortium's criteria (see <http://www.opensource.org/osd.html>). These include the requirement that the source code must be made available so that any user can modify it.

ORDBMS

Object-Relational Database Management System. Also known as an Extensible RDBMS. An ORDBMS is essentially a compromise between the RDBMS and OODBMS models. Data is stored in tables, but there is support for the creation of new and/or complex data types. Oracle, Informix and IBM database products are all ORDBMS's, to some extent, although they do differ in the freedom they allow users to extend the database.

RDBMS

Relational Database Management System. A DBMS which allows data to be stored in multiple related tables. This has several advantages over storing data in a single “flat-file” table, including smaller database size, and better data integrity.

RDF

Resource Description Framework. Provides a way of creating machine-readable metadata to describe resources. RDF is implemented using XML and can use a variety of cataloguing schemes. It also allows for the creation of new cataloguing schemata. RDF is designed to be persistent. As long as a machine reading an RDF description can locate the schema that it uses, it will be able to “understand” the description.

SAX

Simple API for XML. An API for XML that treats the XML document as a stream of events. Unlike the DOM, SAX does not load the whole XML document into memory and may therefore be better at handling very large files.

server-side

Server-side processing means that data processing is done on the server. For example, a script on a Web server might turn the information from a form into a SQL query, send it to a database server, which would run the query and return the information to the Web server. The script on the Web server would then format the results as HTML and deliver it to the client browser.

SGML

Standard Generalized Markup Language. “SGML is a set of rules for defining and expressing the logical structure of documents thereby enabling software products to control the searching, retrieval, and structured display of those documents. The rules are applied in the form of markup (tags) that can be embedded in an electronic document to identify and establish relationships among structural parts.” (EAD Official Web Site, Library of Congress, <http://lcweb.loc.gov/ead/eadback.html>.) HTML is one application of SGML.

SQL

A language designed specifically for working with databases. SQL used to stand for “Structured Query Language,” but it apparently no longer does. It provides functionality such as data retrieval, data manipulation, and defining data structures. Most databases use different “flavors” of SQL, but there are ISO standards for the language. Pronounced “ess-cue-el” by purists, but often pronounced “sequel.”

TEI

Text Encoding Initiative. An implementation of SGML designed specifically for marking up texts in the Humanities.

thin client

In a client/server architecture, a thin client is one that relies on the server to perform most of the data processing.

TIFF

Tagged Image File Format. TIFF is the most widely supported “lossless” image format. This means that it does not discard information when it compresses image files. It is therefore better than JPEG as an archival format, but it also produces much larger files.

XML

Extensible Markup Language. Unlike HTML, which has a predefined set of tags used to mark up text files, XML allows users to create their own tagging schemes. XML tags describe the content of a file rather than its formatting, as HTML does. For example, the title of the *Iliad* would be marked up in HTML as `<I>Iliad</I>`, so that a browser would display it as *Iliad*. In XML, it might be marked `<booktitle>Iliad</booktitle>`. In order for it to display correctly in a browser, the browser would have to understand how to format a book title. This is accomplished using a style sheet (a CSS or XSL file). XML is a fully conforming subset of SGML, so an XML tagging scheme can have a DTD, but it follows stricter rules than SGML in that every tag must be closed (no `<title>` without a `</title>`).

XSL

XML Stylesheet Language. XSL stylesheets are actually XML files themselves. Their function is to provide instructions to a browser telling it how to format an XML file. XSL allows for the transformation of XML documents into other XML documents. For example, a document marked up in TEI XML could be transformed, using XSL, into HTML for presentation on the Web.

APPENDIX C: UNC Digital Library Document Type Definition (DTD)

Proposed Metadata Schema

```
<!--JANUSMETA DOCUMENT TYPE DEFINITION -->
<!--Last modified May 19th, 2000 -->

<!--This DTD is to be used to validate records stored in the UNC Digital -->
<!--Directory database. Records enter this database via a gateway tool -->
<!--named Janus. -->

<!--The DTD is based on the Dublin Core 1.1 recommendations with some -->
<!--modifications by the project staff. See http://purl.org/DC for -->
<!--the Dublin Core 1.1 specification. -->

<!--Comments are prefaced by "#USAGE" followed by the source of the -->
<!--comment if it derives from the DC specification or a DC working -->
<!--group. -->

<!--OBJECT ELEMENT -->
<!--#USAGE:Top level element corresponding to any granularity of aggregation -->
<!--possible in the source collection.-->
<!ELEMENT object (title* , creator* , subject* , description* , publisher* ,
contributor* , date* , type* , format* , identifier* , source* , language* ,
relation* , coverage* , rights* , instance* , meta? )>
<!--#USAGE:Alphanumeric string used to uniquely and permanently identify the -->
<!--object in the context of its source collection.-->
<!ATTLIST object objectid CDATA #REQUIRED>

<!--TITLE ELEMENT -->
<!ELEMENT title (#PCDATA )>
<!ATTLIST title titlequalifier (short |
abbreviation |
alternative |
main |
release |
series |
subtitle |
firstline ) 'main' >

<!--CREATOR ELEMENT -->
<!--#USAGE:DC 1.1: An entity primarily responsible for making the content of -->
<!--the resource.-->
<!ELEMENT creator (#PCDATA )>
<!--#USAGE:DC Agent Qualifier Working Draft 1999-12-10: Indicates the type of -->
<!--the entity for the named Agent.-->
<!ATTLIST creator agenttype (person | organization | event | object )
'person'>
<!--#USAGE:DC Agent Qualifier Working Draft 1999-12-10: Indicates the role of -->
<!--the entity named. They suggest using the list of MARC Relators at -->
<!--http://www.loc.gov/marc/relators/re9802r1.html -->
<!ATTLIST creator agentrole CDATA #IMPLIED>
```

```

<!--SUBJECT ELEMENT -->
<!ELEMENT subject (#PCDATA)*>
<!ATTLIST subject vocabulary CDATA #IMPLIED >

<!--DESCRIPTION ELEMENT -->
<!--#USAGE:DC 1.1: An account of the content of the resource. Description may
include but is not limited to: an abstract, table of contents, reference to a
graphical representation of content or a free-text account of the content.-->
<!ELEMENT description (#PCDATA)>
<!ATTLIST description externaldesc CDATA #IMPLIED
descriptionqualifier (TOC | Abstract) #IMPLIED >

<!--PUBLISHER ELEMENT -->
<!ELEMENT publisher (#PCDATA)>
<!ATTLIST publisher agenttype (person | organization | event | object)
'person'>
<!ATTLIST publisher agentrole CDATA #IMPLIED>

<!--CONTRIBUTOR ELEMENT -->
<!--#USAGE:DC 1.1:Description: An entity responsible for making contributions
to the content of the resource.-->
<!ELEMENT contributor (#PCDATA)>
<!ATTLIST contributor agenttype (person | organization | event | object)
'person'
agentrole CDATA #IMPLIED >

<!--TYPE ELEMENT -->
<!ELEMENT type (#PCDATA)>
<!--#USAGE:DC 1.1:Description: The physical or digital manifestation of the
resource. Typically, Format may include the media-type or dimensions of the
resource. Format may be used to determine the software, hardware or other
equipment needed to display or operate the resource. Examples of dimensions
include size and duration. Recommended best practice is to select a value from
a controlled vocabulary (for example, the list of Internet Media Types [MIME]
defining computer media formats).-->
<!ELEMENT format (#PCDATA)>
<!ATTLIST format formatscheme (mime | other) 'mime'
formatresolution (high | medium | low) #IMPLIED
filesize CDATA #IMPLIED >

```

```

<!--IDENTIFIER ELEMENT -->
<!--#USAGE:DC 1.1:Description: An unambiguous reference to the resource within
a given context. Recommended best practice is to identify the resource by means
of a string or number conforming to a formal identification system. Example
formal identification systems include the Uniform Resource Identifier (URI)
(including the Uniform Resource Locator (URL)), the Digital Object Identifier
(DOI) and the International Standard Book Number (ISBN). NB: in our scheme,
this is distinct from the object's ID in the source system.-->
<!ELEMENT identifier (#PCDATA )>
<!ATTLIST identifier identifierscheme (DOI |
                                         ISBN |
                                         ISMN |
                                         ISRC |
                                         ISRN |
                                         ISSN |
                                         SICI |
                                         URL |
                                         URN |
                                         collectionspecific |
                                         callno ) 'URL'
        identifierid ID #IMPLIED >

<!--SOURCE ELEMENT -->
<!--#USAGE:DC 1.1:Description: A Reference to a resource from which the present
resource is derived.-->
<!ELEMENT source (#PCDATA )>

<!--LANGUAGE ELEMENT -->
<!--#USAGE:DC 1.1:Description: A language of the intellectual content of the
resource.-->
<!ELEMENT language (#PCDATA )>

<!--RELATION ELEMENT -->
<!--#USAGE:DC 1.1:Description: A reference to a related resource.-->
<!--#USAGE:Comment: For the prototype, this is used to indicate the primary key
of the collection to which the object belongs (if there is one).-->
<!ELEMENT relation (#PCDATA )>
<!ATTLIST relation relationtype (PartOf | References ) 'PartOf' >

<!--COVERAGE ELEMENT -->
<!--#USAGE:DC 1.1:Description: The extent or scope of the content of the
resource. Coverage will typically include spatial location (a place name or
geographic coordinates), temporal period (a period label, date, or date range)
or jurisdiction (such as a named administrative entity).-->
<!ELEMENT coverage (#PCDATA | coordinate )*>
<!ATTLIST coverage coveragetype (placename |
                                   periodname |
                                   stylename |
                                   time |
                                   point |
                                   line |
                                   polygon ) 'placename'
        coveragescheme CDATA #IMPLIED
        coveredatum CDATA #IMPLIED >

```

```

<!--COORDINATE ELEMENT -->
<!ELEMENT coordinate EMPTY>
<!ATTLIST coordinate coordinateorder CDATA '1'
                coordinateaccuracy (exact | approximate ) 'approximate'
                coordinatetype (x |
                                xmin |
                                xmax |
                                y |
                                ymin |
                                ymax |
                                z |
                                zmin |
                                zmax |
                                t |
                                tmin |
                                tmax ) #IMPLIED
                coordinateeval CDATA #REQUIRED >

<!--RIGHTS ELEMENT -->
<!--#USAGE:Comment: Rights and Identifiers are paired somehow for permissions
for each proxy. Identifiers will have numerical (incremental) id numbers, which
the rights entries will reference. The pairing of rights with proxies is
accomplished by means of the instance tag--one rights tag per instance.-->
<!ELEMENT rights (rightsstatement , rightscontact* , rightsrealm+ )>

<!--RIGHTSCONTACT ELEMENT -->
<!ELEMENT rightscontact (#PCDATA )>
<!ATTLIST rightscontact contacttype (URL | email | freetext ) 'freetext' >

<!--RIGHTSREALM ELEMENT -->
<!--#USAGE:Comment: The rightsrealm is a reference to one or more entities in
the User Management DB (Adso). The text of the element is the entity name, and
the 'entityid' attribute is its key in the database. The rightsrealm that
allows global access is 'other'.-->
<!ELEMENT rightsrealm (#PCDATA )>
<!ATTLIST rightsrealm discover CDATA 'true'
                read CDATA 'true'
                write CDATA 'false'
                execute CDATA 'false'
                entityid CDATA #REQUIRED >

<!--RIGHTSSTATEMENT ELEMENT-->
<!ELEMENT rightsstatement (#PCDATA )>

<!--DATE ELEMENT -->
<!--#USAGE:Comment: Dates apply to the object referenced by the DC metadata.
The original may or may not be digital in nature. Date information related to
the item depicted belongs in Coverage. -->
<!ELEMENT date (#PCDATA )>
<!ATTLIST date datatype (Created |
                        DataGathered |
                        Valid |
                        Issued |
                        Available |
                        Accepted |
                        Acquired ) #IMPLIED
                dateaccuracy (exact | approximate ) 'exact' >

```

```
<!--META ELEMENT -->
<!--#USAGE:Comment: Not a Dublin Core tag!  Meta allows for the storage of
'extra' metadata which can't easily be parsed into the Dublin Core format.
Eventually, we anticipate being able to use this for customizable searches on
individual projects' resources.-->
<!--#USAGE:Comment: The DTD will have to be extended to handle information
stored in <meta>.  In effect, we will have a DTD for each possible variation.
XML namespaces may provide a solution to this problem. -->
<!ELEMENT meta ANY>
```

Sample XML

The following file was converted from the Digital Scriptorium's collection of Historical American Sheet Music. The original metadata was encoded in EAD (Encoded Archival Description) SGML.³⁹

```
<?xml version="1.0"?>
<!DOCTYPE object SYSTEM
"http://adso.classics.unc.edu/shadow/janus/janusmeta.dtd">
<!--Janus conversion of preprocessed Digital Scriptorium American Sheet Music
source to Janus-generic metadata-->
<!--janus_DS_ASM.xsl version date: 26 March 2000-->
<!--Converted on: 4/29/00 2:06:03 PM-->
<object objectid="DS-HASM00a6448">
  <title titlequalifier="main">Gen. Quitman's grand march</title>
<creator agenttype="person" agentrole="Composer">Eaton, E. O. (Edward O.), 19th
cent.</creator>
<creator agenttype="person" agentrole="Lithographer">Schnabel &
Finkeldey</creator>
  <subject vocabulary="HASM Subject Content">Marches and Military
Music</subject>
  <subject vocabulary="HASM Illustration Type">Personalities--Quitman, John
A.</subject>
  <subject vocabulary="LCTGM">Gen. John A. Quitman</subject>
  <subject vocabulary="LCTGM">Governors</subject>
  <subject vocabulary="AAT">governors</subject>
  <subject vocabulary="LCSH">Marches (Piano)</subject>
  <subject vocabulary="LCSH">Quitman, John A.--Pictorial works</subject>
  <description>Pagination: 5. Instrumentation: piano. Dedicated to Gen. John
A. Quitman.</description>
  <publisher agenttype="organization" agentrole="Publisher">Beck &
Lawton, Philadelphia, Pennsylvania.</publisher>
  <date datatype="Issued" dateaccuracy="exact">1858</date>
  <type>sheet music</type>
  <identifier identifierscheme="callno">Music A-6448</identifier>
  <identifier identifierscheme="collectionspecific">a6448</identifier>
  <language>en-us</language>
  <relation relationtype="PartOf">DS-HASM0000001</relation>
  <rights>
    <rightsstatement>&copy;1998 Duke University. All rights reserved. Images
and texts on these pages are intended for research and educational use
only. Please read our statement on use and reproduction at
http://scriptorium.lib.duke.edu/specoll/copyright.html for further
information on how to receive permission to reproduce an item or how to
cite it.</rightsstatement>
    <rightscontact contacttype="URL">
http://scriptorium.lib.duke.edu/specoll/copyright.html</rightscontact>
    <rightsrealm discover="true" read="true" write="true" execute="true"
entityid="1">Digital Scriptorium</rightsrealm>
    <rightsrealm discover="true" read="true" write="false" execute="true"
entityid="0">other</rightsrealm>
  </rights>
</object>
```

³⁹ The document from which the example was abstracted can be found at <http://scriptorium.lib.duke.edu/sheetmusic/1850-1859.txt>.

```

<instance>
  <creator agenttype="person" agentrole="document encoding">Sheet Music
  database prepared by Lois Schultz and encoded by Stephen Douglas Miller,
  Rare Book, Manuscript, and Special Collections Library, Duke University.
  October 1998</creator>
  <publisher agenttype="organization" agentrole="Publisher">Rare Book,
  Manuscript, and Special Collections Library Duke University</publisher>
  <date datatype="Issued" dateaccuracy="exact">1998-10</date>
  <type>image</type>
  <format formatscheme="mime">text/html</format>
  <identifier identifierscheme="URL">
  http://scriptorium.lib.duke.edu/sheetmusic/a/a64/a6448</identifier>
  <rights>
    <rightsstatement>&copy; 1998 Duke University. All rights reserved.
    Images and texts on these pages are intended for research and
    educational use only. Please read our statement on use and
    reproduction at http://scriptorium.lib.duke.edu/specoll/copyright.html
    for further information on how to receive permission to reproduce an
    item or how to cite it.</rightsstatement>
    <rightscontact contacttype="URL">
    http://scriptorium.lib.duke.edu/specoll/copyright.html</rightscontact>
    <rightsrealm discover="true" read="true" write="true" execute="true"
    entityid="1">Digital Scriptorium</rightsrealm>
    <rightsrealm discover="true" read="true" write="false" execute="true"
    entityid="0">other</rightsrealm>
  </rights>
</instance>
</object>

```

APPENDIX D: Further Information

The following list of Internet resources represents a limited selection of materials useful for understanding the impetus, challenges, and goals of the UNC Digital Library Project. For a more extensive list of such Internet resources, please see <http://www.unc.edu/campus/its/projects/diglib/links.htm>.

Internal UNC Digital Library Documents

- Oracle *interMedia* Checklist
http://www.unc.edu/campus/its/projects/diglib/Vendor_Checklist-Oracle.html
- Informix Media360 Checklist
http://www.unc.edu/campus/its/projects/diglib/Vendor_Checklist-Informix.html
- Cross-reference Checklist (*interMedia* vs. Media360)
<http://apollo.classics.unc.edu/IIMD/checklist.asp>

Oracle and Informix White Papers

- Oracle *interMedia* White Papers (General)
<http://www.oracle.com/o8i/owa/O8i.Search?string=intermedia>
- Oracle *interMedia* White Paper (Managing Multimedia Content)
http://www.oracle.com/database/documents/managing_content_twp.pdf
- Informix Multimedia White Papers (General)
<http://www.informix.com/informix/whitepapers/>
- Informix Media360 White Paper
<http://www.informix.com/informix/whitepapers/aberdeen/media360.htm>

Electronic Publishing and Digital Libraries

- American Memory: <http://memory.loc.gov/>
- The Digital Library Federation: <http://www.clir.org/diglib/dlhomepage.htm>
- The Stoa: <http://www.stoa.org/>
- Institute for Advanced Technology in the Humanities: <http://www.iath.virginia.edu/>
- Museum Educational Site Licensing Project (Final Report, July 1998):
<http://sunsite.berkeley.edu/Imaging/Databases/1998mellon/toc.html>
- The Oxford Text Archive: <http://ota.ahds.ac.uk/>
- The Visual Resources Association: <http://www.oberlin.edu/~art/vra/dsc.html>
- Scoping the Future of the University of Oxford's Digital Library Collections (Final Report):
<http://www.bodley.ox.ac.uk/scoping/report.html>