

Digitization and Metadata:

Independent Study Final Project

David Gwynn

LIS 690 - Dr. Nora Bird

4 May 2009

INTRODUCTION

For each module, there are general comments, key concepts (as I see them), and notes on the readings. Readings from the modules are included in the order they appeared within the module. I have included comments on all the required readings and on ant optional readings that I felt were worthy of attention or addition to the list. I have also included a few additional readings I think may be of interest. There is also a table of all readings at the end of the document.

MODULE 1-B: HISTORY OF DIGITAL LIBRARIES AND LIBRARY AUTOMATION

General Thoughts

While many of the readings for this module are useful and interesting, there should also be a sort of unifying overview that might tie everything together. This could be found in the textbook for the course. This module might also be a good place for a discussion of different digital projects as a means of determining just what a digital library is. The term “digital library” so far has not really been satisfactorily defined here, except for some good discussion in Greenstein & Thorin. Focusing on digital projects would also tie this module in with the next one.

Key Concepts

- Definition and focus of “digital library”: reformatting and digitization, databases and third-party digital resources, or just “born digital” materials?
- Evolution of the concept from systems/internal focus (OPAC) to patron focus (CD-ROM databases) to public access (web access).
- Digital libraries as distinct units or as part of the existing library.
- End user expectations.

Required Readings

Wattenberg (1998)

While this article has considerable historical interest for advanced students and for those who are really interested in the background of digital libraries (and of certain

internet projects in general), it doesn't offer a lot of really useful information for the beginning or mid-level student, particularly given that most of the illustrative external links no longer work.

Griffin (1998)

This article gives useful and brief background information on the Digital Libraries Initiative, and is probably worth reading for most students who are interested in the subject of digital libraries.

Greenstein & Thorin (2002)

This article is definitely useful at all levels because of its emphasis on comparing how different institutions (all of them academically-based, which may be a weakness) have both defined and implemented the concept of a digital library. The discussions of funding, the experimental (as opposed to fully functional and useful) nature of early initiatives, and respect for standards and interoperability were particularly interesting, and the latter will be a recurring theme throughout the rest of the modules as well. Greenstein & Thorin also exhibit some concern with the experience and expectations of end users which is missing in some of the later modules and readings.

Other Useful Readings from Module

Bush (1945)

This article should probably be required reading for anyone who considers himself a student of information management. In addition to presenting an interesting view of then-current advances in information technology, Bush also predicts such future issues as preservation of electronic material, spreadsheet and database software, “real” costs of content creation, and compression and storage issues.

Section F: Major Digital Library Projects

This list of projects might be useful to present to students as a list of sites to scan, or perhaps as references and visual aids during a lecture. Following are some other interesting projects, several of which are not mentioned in the module, that might be worth a look:

- SF Public Library Digital Photograph Collection: <http://sflib1.sfpl.org:82/search>
- California Digital Library: <http://www.cdlib.org/>
- Digital Pittsburgh: <http://digital.library.pitt.edu/pittsburgh/>
- David Rumsey Map Collection: <http://www.davidrumsey.com/>
- JSTOR: <http://www.jstor.org/>

MODULE 3-B: DIGITIZATION

General Thoughts

The digitization module seemed one of the best developed and easiest to follow for me, but that is likely due to the fact that it's my area of interest, and I'm already familiar with most of the issues, theory and terminology.

Digitization and metadata go hand in hand; it's really difficult to discuss one without the other; therefore, many aspects of this module and the following one overlap in significant ways. Digitization mechanics must take metadata into consideration, while many aspects of metadata are dependent and contingent upon digitization formats, server configurations, etc. Particularly at the level of a collection management class, it might be easier to consider the two as one unit, because digitization is not purely about the mechanics.

Key Concepts

- Planning and selection.
- Copyright.
- Curated exhibit vs. mass project.
- Preservation vs. access.
- Description and metadata (next module).
- Formats and delivery.

Required Readings from Module

Chowdhury & Chowdhury (2003)

This is a good overview of terms related to digitization, introduction to formats, etc., and is more practical than theoretical.

Cornell University Library (2000)

This is absolutely one of the most useful items in the whole module. I completed it in another class (LIS 505) as well, and it gave me a great understanding of the material. It could also be used as part of a lecture.

Smith (1999)

This is a good piece that makes clear the distinction between digitization and preservation, the potential for questions of provenance with digital objects (and comment on how we rely on repositories as proof of authenticity), hybrid means of preservation (scan to microfilm, etc.), digital surrogates as a means of preservation, curated collections as “publications”, user expectations, ethics and issues of selection, and copyright. The article also notes the “added value” nature of digital collections

Recommended Readings from Module

Liu (2004)

This is a good overview of theoretical issues and concerns, as opposed to Chowdhury and Chowdhury’s more practical treatment above. I also reviewed this in a

bibliography I created in 2007, saying, "This survey of digitization projects, primarily among academic libraries in the United States, attempts to determine standards for digitization projects and determine types of materials and issues surrounding digitization. It also discusses special challenges at the Internet Archive, a project that involves archiving of websites ("The Wayback Machine") and other digital material, including audio and video."

Humanities Advanced Technology and Information Institute (2002)

This article is a good take on selection criteria. Some of the material may seem a bit repetitive given the other readings, but I think there's enough original material to make it worthwhile to include, particularly if this is to be a component in a collection management class.

Hazen & Merrill-Oldham (1998)

Given its age, and the fact that it covers so much material covered elsewhere (particularly Humanities Advanced Technology above), this article could easily be eliminated.

Other Useful Readings from Module

Wisser (2007)

The NC ECHO guidelines would be worth a look by advanced students or those who wanted to go into a little deeper detail. Wisser teaches courses on archival metadata

for the Society of American Archivists, and also is very conversant in EAD (Encoded Archival Description) and cataloguing theory.

Peterson (2004)

This is an excellent guide to technical specifications for photo scanning, which is (and probably will continue to be) the most prominent type of digitization project. This would be good for advanced students.

Additional Possibilities Not Included in Module

Hughes (2004). *Digitizing collections : strategic issues for the information manager*. London: Facet.

This book is a management handbook for digitization projects and contains individual chapters on costs and benefits, material selection, legal issues, planning and funding, project management, and material types, some of which might make interesting additional readings. This would be most useful for moderate to advanced students.

Koelling (2004). *Digital imaging: a practical approach*. Walnut Creek, CA: Altamira Press.

This is a handbook on digital imaging and is one of the best I have seen. In addition to mechanics, it also touches on copyright and ethical issues as well as forensic research. This would probably be more useful for advanced students who are planning to concentrate on photo/visual projects.

MODULE 4-B: METADATA

General Thoughts

Many students seem to be very confused about the concept of metadata, and don't quite understand how it's used, and in particular seem to be baffled by XML and how it relates to different formats and schemas with which they may already be familiar, such as MARC and Dublin Core.

It might be really helpful for some students to see how this works with applications they use on a regular basis (e.g. iTunes, Adobe Bridge, RSS feeds for websites, etc.), particularly with respect to the idea of embedded metadata vs. metadata stored in a separate file. A JPEG or MP3 file is capable of holding certain metadata on its own, as part of the specification of the file format (length, encoding info, thumbnail, etc.), but iTunes and other programs also add certain contextual metadata (like playlist, genre, etc.) Introduction of DACS (*Describing Archives: A Content Standard*) might be beneficial at this point, along with its relationship to EAD, since most digitization projects also have something of an archival basis. There could also be more information on things like metadata crosswalks and interoperability (although this is discussed to some extent in the architecture module), MARC, etc.

In my opinion, this is the kind of thing that should be taught in a modern cataloguing class.

Key Concepts

- Items in differing metadata schemas (MARC, DC, DACS).
- Interoperability and crosswalks.
- Relationship between XML and metadata.
- Input/output formats.

Required Readings from Module

Weibel (1995)

This is very interesting as a report on the development of Dublin Core and the rationale behind it (simplicity, expandability, etc.), but it's not really full of essential information. It could probably be condensed into a table or a Powerpoint slide.

Duval, Hodgins, Sutton, & Weibel (2002)

This is a generally good look at the theoretical underpinnings of metadata theory and is particularly useful for description of embedded, associated, and third party metadata, syntax vs. semantics, mandatory vs. optional fields, and subjective vs. objective description (including the contextual aspects of each).

Other Useful Readings from Module

Dublin Core Metadata Initiative.

This would be for more advanced students who wanted to learn more, or perhaps for students in a metadata-specific or cataloguing class. It might be useful for all students to look at some sample Dublin Core output as a visual aid, though.

The Open Archives Initiative Protocol for Metadata Harvesting.

This could be useful for more advanced or specific students, but it is definitely a bit heavy of the technical information. The same could be said for all the information on preservation metadata (Section G).

Additional Possibilities Not Included in Module

National Information Standards Organization (2004). *Understanding metadata*.

Bethesda, MD: NISO Press

This should be the starting point for this section even though it's not part of the module. It features an excellent overview of the concept and application of metadata in easy to understand language.

North Carolina ECHO (2007). North Carolina Dublin Core template. Retrieved April 25, 2009, from NC ECHO Web site:

<http://www.ncecho.org/dig/ncdctemplate.shtml>

This online application generates metadata compatible with the North Carolina ECHO guidelines and based on Dublin Core. It could be a useful hands-on demonstration tool and might also be useful for student projects.

MODULE: 5-A: DL ARCHITECTURES

General Thoughts

There were, in my view, some major problems with this whole module. There was no overview and I was not really certain what the aim of the module was, and how or why it was distinct from the software module. For students who are not concentrating in information technology or computer science, there seemed to be very little that was of much use here, and I sense that most LIS students will neither understand, care about, nor retain very much of this material. Students need to be somewhat fluent in “tech speak” so they can ask for what they need and converse intelligently with IT people, but most of this material seemed to go a few steps too far. It was just too detailed. I don’t think detailed technical training is what these courses would really be all about.

Ultimately, I feel this entire module could probably be skipped for students not in a semester-long class, although a few of the key concepts might be incorporated into a lecture.

Key Concepts

- Creating digital object identifiers as distinct from URLs.
- Federated vs. union search.
- Interoperability.

Required Readings from Module

Arms (1995)

This is an interesting article for historical overview, as it discusses early theories on digital object identifiers in terms of the then-developing DNS/URL protocols. The article also mentions the RDA (object and manifestation) concept. As a side note, an interesting comment on page 8 almost predicts what would later become the PDF format.

Kahn & Wilensky (1995)

This is for advanced or very interested students only. Many of the key concepts are also covered in Arms above.

Gonçalves, France, & Fox. (2001)

This includes a good description of the benefits of federated vs. union search, but otherwise it seems much more detailed than necessary.

McNab, Witten, & Boddie (1997)

This article was a bit overly technical. One issue was that the authors kept referring to “documents” rather than “objects” or even “items”. While this is standard computer operating system terminology, it still speaks to a text-based bias and to the absence of end user focus; most users do not think of a piece of music (or a photo or video), no matter how it’s represented, as a “document”.

Petinot, Giles, Bhatnagar, Teregowda, Han, & Council (2004)

I didn't find this terribly helpful at all. Perhaps it was too technical in nature.

Suleman & Fox (2001)

This article is worth a look. It covers general concepts well, possibly better than above.

MODULE: 5-B: APPLICATION SOFTWARE

General Thoughts

This section could also use an overview, and the whole module is probably extraneous for students not taking a semester-long course. It seems too much to expect that students will understand application software just through comparisons of specific packages without a “big picture” of what the whole process entails, which is, in a nutshell, the process of making a centralized database and digital objects interact with an end user’s web browser. There’s very little discussion of such topics as PHP, MySQL, and ASP in general, only profiles of specific software packages. There was not enough general information on software basics, and maybe too much information on the specifics of each package. These are all essentially content management systems; the only difference is the variables added to the database for each item, and how interoperable those databases are. The information here in some ways actually makes it much more complicated than it really needs to be.

Again, most of this comes down to working with tools to create interactive database-driven websites, and that should be made a little clearer than it is right now.

Extensive discussion of various packages will be interesting to more advanced students, but is far too detailed for most students in a collection management class.

It may be appropriate for those in a standalone class, but perhaps in a slightly condensed format.

With the evaluations of Eprints and other software packages, there is considerable discussion of the interface's ease of use for users who are *submitting* content, but not nearly as much about the (more important) users who will be *accessing* that content. Only the CONTENTdm material really addresses that satisfactorily.

Key Concepts

- What is a content management system?
- Platform (PHP, ASP, etc.)
- Open source vs. commercial product.
- Input/output formats.
- End user experience.

Required Readings from Module

Eprints, DSpace, Greenstone, and CONTENTdm Software Manuals and Introductions

These are interesting documents for more advanced students to scan or read, but a condensed table or Powerpoint slide would probably provide enough perspective for students in a content management class.

Suleman & Fox (2001)

This article, also included in Module 5-A is worth a look. It covers general concepts well, possibly better than above. The CONTENTdm articles might bear closer scrutiny because it is something of an "industry standard" and also because it is the only one that really gives detailed attention to the end user experience. Also, most of the

comparative articles that follow focus only on the open source options like Greenstone and Eprints.

Witten & Bainbridge (2005)

This was a very interesting document for me, because I have some interest in specifics of database interoperability and was able to follow it to some extent. I can't imagine, though, that it would be of any significant interest to most students taking a digital libraries course. It's just too technical and detailed.

Don, Bainbridge, & Witten (2005)

Again, this is interesting but probably too detailed for most students.

Wang, Assion, & Matthaei (2003)

This article is useful for its comparison of Eprints and Greenstone. It might be worthwhile to assign the first few pages, letting more advanced students skim the rest.

Goh, Chua, Khoo, Khoo, Mak, & Ng (2006).

This is helpful for comparison of the various open source packages. Like the others, this is probably better for somewhat more advanced students.

MODULE 9-C: DIGITAL LIBRARY EVALUATION, USER STUDIES

General Thoughts

Library service evaluation could very easily be a semester-long class all by itself. Evaluation of digital libraries would be an important component of a standalone DL course or of a DL section in a collection management course, particularly since the end user experience is not well covered in the other modules. Site usability testing is covered in two web design courses within the LIS department, so testing of the technical aspect of DLs might best be covered there. In a collection management or standalone class, it might be better to focus on the content aspects rather than specifically on usability.

Key Concepts

- Approaches: bibliometric research, log analysis, observational, user surveys.
- Necessity for and definitions of multiple collection methods (purely technological methods do not always reflect patron intent).
- User-based focus.
- Proactive vs. reactive evaluation.

Required Readings from Module

Nicholson (2004)

This article presents a general overview of library evaluation theory and practice, and is not necessarily specific to digital libraries. This would probably be a useful addition

to a collection management class (or perhaps even a information literacy class) with or without a DL component. It is significant that the article discusses digital services within the context of other library services rather than in isolation.

Reeves, Apedoe, & Hee Woo (2003)

Presented in more of a textbook format, these two chapters carry forward the theory behind different methods of evaluation into a more specifically digital program. This is an introduction to methods discussed at more length in subsequent chapters.

Students who are more advanced and interested might benefit from some of the more detailed material as well, particularly the chapter on reporting.

Other Useful Readings from Module

Most of the additional readings expand on the material covered in the required readings, with some taking on one specific aspect, such as bibliometric studies (Bollen & Luce, 2002). There continues to be a bit of an issue with the definition of a digital library; Bryan-Kinns & Blanford (2000) assume the term refers purely to article databases and full-text searches, but the article could still be applicable to other types of digital libraries. Chowdhury, Hobbs, & Flores (2002) and Reiger (1999) discuss issues related to digital image libraries. Saracevic (2005) offers something of an outline of approaches to evaluation. Thong (2002) focuses more on the user interface issues that might better be discussed in a web design class, as mentioned above. While interesting, these would probably best be considered optional readings for advanced students.

