
This study describes the results of an assessment of the current state of digital library evaluation. Representatives from eleven digital library teams that have conducted an assessment were interviewed to determine what motivates an evaluation and how it is defined and carried out. The study also examines these evaluation case studies for characteristics of longevity, adaptability, and triangulation.

Identifying the user population and their information needs were most commonly cited as the reasons for undertaking an evaluation. Objectives ranged from improved interface and information retrieval functions to collection management. Methodologies included online surveys, transaction log analysis, and usability testing. Only two of the eleven digital library projects interviewed have conducted long term evaluations. Several of the projects displayed characteristics of adaptiveness in that lessons learned from one evaluation were applied to other digital library assessments. Half of the four evaluations teams that used more than one methodology analyzed the resulting data using triangulation.

Headings:

Virtual library -- Evaluation

Information systems -- Evaluation

Research techniques

Use Studies – Virtual library
Introduction

Digital libraries are a rapidly evolving technology growing out of interdisciplinary research. Librarians, information scientists, computer scientists, and researchers in other fields bring multiple perspectives to digital library development, evidenced by the diversity of digital library projects in existence today. Researchers and digital library teams have focused on problems of information retrieval, metadata, digital objects, collection development, and information acquisition and use (Borgman, 2000). Identifying and standardizing the best solutions to these problems will require systematic assessment. However, evaluation has not become an integral part of digital library development.

So far, evaluation has not kept pace with efforts in digital libraries (or digital libraries themselves), has not become a part of their integral activity, and has not been even specified as to what it means and how to do it (Saracevic, 2000, p. 351).

Recognizing the need for systematic evaluation, the digital library community is beginning to discuss how to define and carry out assessments. Long-standing digital library projects—such as Alexandria Digital Library, Making of America and Perseus Digital Library—have published articles that reflect upon their evaluative research (Bishop et al., 2000; Kilker and Gay, 1998; Marchionini, 2001). Several common themes emerge from these reports. Digital library evaluations should persist over time and adapt as new insights emerge. Digital library assessment is exploratory, which requires evaluators to collect data from a variety of measures to identify new patterns and relationships.
This study investigates the current state of digital library evaluation. Eleven digital library teams who have conducted an assessment were interviewed to determine what motivates an evaluation and how is it defined and executed in practice. The study also examines these evaluation case studies for characteristics of longevity, adaptability, and triangulation.
Literature Review

Researchers have offered several theoretical frameworks to describe the nature of digital libraries as emerging phenomena. The team that developed DeLIver, a digital library of full-text computer science and physics journals, describes the development of digital libraries as assemblages, a dynamic created by the interaction between people, their ideas, beliefs, and practices about technology, and the technology itself (Bishop et al., 2000). The concept of technology as an assemblage views development as a negotiation process over time.

A related framework, the social construction of technology (SCOT) theory, describes technology development as a process of multiple negotiations between various social groups and the developing technology. The theory is based on three main concepts: relevant social groups, interpretive flexibility and closure. SCOT suggests that a single technology may develop “multiple branches” to meet the needs of various groups—relevant social groups. Proponents of this theory study how these groups’ varying concepts of the technology—interpretive flexibility—lead to a stabilized adaptation of that technology—closure (Kilker and Gay, 1998). Kilker and Gay applied the social construction of technology theory to describe the development and evaluation of the Making of America digital library.

Understanding the interactions and adaptations between users, environment, and technology is the key to guiding the emergence of digital libraries as a technology. This
requires formative evaluation, which “aims to understand the meaning of some phenomenon situated in a context and the changes that take place as the phenomenon and the context interact” (Marchionini, 2001, p. 311). Formative evaluation is exploratory and necessitates the objectives and methodologies of the assessment adapt as the evaluator’s understanding of digital libraries grows.

Meeting the challenge of “assessing complex adaptive systems” (Marchionini, 2001, p. 305) involves more than just collecting data. An effective evaluation operationalizes the construct considered for assessment, the context in which it is evaluated, the performance criteria, measures, and methodologies (Saracevic, 2000).

Operationalizing the construct under evaluation, in the broadest sense, requires the evaluator to define the term ‘digital library’. This is no small task; the literature abounds with competing definitions and the debate continues. However in any particular instance of an assessment, it is enough to define which elements of a digital library will be evaluated. Evaluators may choose to focus on the interface or the search function or some type of service provided by the digital library.

Once the construct is defined, the context in which the evaluation will take place must be determined. Choosing a context is like selecting a lens through which the digital library will be viewed. The context may be user-centered or system-centered. User-centered evaluations may study the interaction between a single individual and the system. Evaluators can take a broader view of the user as an institution or community, in which case the focus is on how the digital library meets the needs of the larger group (Saracevic, 2000). Evaluations focusing on the individual are more common, probably because the interactions are easier to isolate and measure; however, the results of this
type of study cannot be generalized to larger groups (Saracevic, 2000). System-centered evaluations are easier to quantify and replicate and fall into three categories: engineering, processing, and content. An evaluation focused on engineering considers how well the hardware and networks perform. At the processing level, performance of procedures, techniques, algorithms, and operations are measured. Finally the digital library can be assessed in terms of content selection, representation, and organization (Saracevic, 364). In reality these user groups (individual, institution, community) and aspects of the system (engineering, processing, content) are interdependent entities, but they have not yet been evaluated simultaneously. This is a shortcoming in the evaluation process (Saracevic, 364).

The next consideration is to determine what criteria will be used to describe the system’s performance in relation to the established context. Digital libraries have evolved out of several long established disciplines: library and information sciences, human computer interaction, and computer science. Criteria from these fields can be used to assess digital libraries. Libraries have always assessed their collections in terms of purpose, subject, scope, authority, coverage, currency, audience, cost, format, treatment, preservation, and persistence, and judge individual information sources based on their accuracy, appropriateness, representation, uniqueness, comparability, presentation, timeliness, and ownership. As a type of information retrieval system, digital libraries can be assessed using measures of relevance and satisfaction. Criteria from the study of human-computer interaction and interface design, such as usability, task appropriateness, design features, navigation, and browsing, are some of the most commonly applied measures (Saracevic, 2000).
Once the criteria have been selected, evaluators must decide how to measure the performance of those criteria. This has proven to be problematic, because there are no agreed upon standards for measuring even the most fundamental aspects of performance. For example, evaluators often use statistics generated from transaction logs to measure digital library use. However they may report usage as a raw number of “hits”, the number of sessions, or by the number of unique IP addresses. Many measures are too ambiguous to interpret with confidence. Can one say that a “hit” constitutes use? Even though measures can be difficult to determine and harder to analyze, the digital library community will benefit from an effort to standardize measures (Bishop, 2000).

Finally, evaluators must select a methodology that best matches the context, objectives, and performance criteria. Methodologies outline how data will be collected and who will be included in the assessment. Evaluators can collect data by observing the user or by gathering users’ opinions about the digital library. Data collection by observation may be accomplished with assisting technologies, such as tracking software or audio-visual recorders. Users’ opinions are typically gathered through surveys and interviews. Data can be collected synchronously or asynchronously. Synchronous methods, such as think-aloud sessions, provide immediate and spontaneous feedback. Asynchronous data collection methods more closely model a natural environment and permit time for reflection (Reiger and Gay, 1999).

Surveys, transaction log analysis, and usability testing are some of the most common methodologies employed for digital library evaluation. Surveys can be useful instruments when the questions have been carefully designed. Generally, they do not require a great deal of resources or coordination to implement and are versatile enough to
capture everything from basic demographic information to rich qualitative data. However, surveys have their disadvantages. Individuals’ responses may not generalize to the population as a whole and may not be completely honest, if they are concerned about their privacy (Parsons, 2000).

Transaction logs monitor user actions by tracking clicks, keystrokes, items viewed, and navigation strategies. Evaluators typically use transaction logs to measure use, search outcomes, time, and study navigation patterns and identify common errors that indicate poorly designed features (Borgman et al., 1998). Logging software often comes loaded on web servers or can be obtained open source. Evaluators who find commercial software to be inadequate, design their own customized transaction log tools.

Usability covers a broad range of user-centered methodologies designed to assess a system’s ease of operation. Usability methodologies can be applied to improve a system under development (formative) or to describe the effectiveness of the completed site (summative). In the earliest stages of development, usability methods—such as ethnographic evaluation and task analysis—collect users’ system requirements through observation and interviews. In a heuristic evaluation, users or user surrogates (inspectors) assess the system according to predefined principles. These principles may include: user control, consistency, error prevention, recognition rather than recall, aesthetic and minimalist design, and documentation (Nielsen, 1994). Cognitive walkthroughs involve experts performing typical tasks with a prototype and analyzing the system’s ability to support those tasks effectively. After a working prototype has been developed, evaluators can conduct usability tests using the think-aloud method in which users describe their reactions and explain their actions while they are working with the
system. Such usability testing yields rich and useful data, but it has its disadvantages. These methodologies can require more staff, a greater time commitment on the part of the users and the evaluators, and special equipment or testing environment. Some types of usability testing, such as the think-aloud method, may be awkward for the user and create an artificial environment.

An investigation that addresses these five components—construct, context, performance criteria, measures, and methodologies—meets the basic requirements of an evaluation, but is not sufficient to provide insight and guide the development of digital libraries as an emerging technology. Evaluation should be a long term, adaptive process that looks for meaning through multiple measures.

Assessments should be a process over time, since digital libraries evolve as the users and technology “mutually adapt and mature” (Marchionini, 2000, p. 305). Marchionini (2000) illustrates this point with a case study of the Perseus Digital Library. Problems with hardware and software were a major issue in the first four years of the project. As delivery mechanisms improved and users became more comfortable with the technology, problems with the “physical infrastructure” no longer dominated assessment results. If evaluation had ended in the early years of the project, temporary circumstances would have skewed the findings.

Evaluations must be flexible to accommodate changes that occur in the technology over time. Kilker and Gay (1998) describe how the assessment of the Making of America digital library was an adaptive process: “... [T]he MOA project used evaluation within an iterative design context. The evaluation was therefore exploratory: although guided by general goals, we did not know how people would use the
technology, or what aspects of the technology would be the greatest barriers. The methodology of the evaluation itself was revised in an iterative fashion” (68). Program evaluation research refers to this as an active-reactive-adaptive approach. Evaluators actively design an assessment based on the users and context, react to new information generated by the assessment, and adapt the evaluation accordingly (Williams, 2002).

Finally, digital library evaluation is an exploratory process. Evaluators should employ triangulation, an analysis technique which compares various information sources to uncover new relationships. Triangulation occurs when the evaluator collects multiple data sets or uses several different methodologies.
Research Questions

While the literature reflects a growing interest, evaluation has not become an integral part of digital library development and maintenance. As digital library projects proliferate and mature, how does this affect the state of evaluative efforts? This study will survey 200 digital libraries and interview representatives from projects that have conducted evaluations to determine:

Question 1: What are the motivations for assessment?

Question 2: How is evaluation defined and executed in practice? What are the objectives? Which methodologies are used?

Question 3: Do practical assessments exhibit characteristics of a model evaluation as defined in the literature? Are they long-term, adaptive, and do they incorporate multiple measures?
Methodology

An email survey was administered to 200 digital libraries in order to identify projects that had been evaluated. Team members from digital libraries that had been assessed were interviewed to determine the motivations, methodologies, and outcomes of digital library evaluations. Details of the methods used in this two-phase study are provided in this section.

Survey Sample

The majority of digital libraries surveyed were selected from digital initiatives listings found on professional organizations’ websites and online registries. The listings were browsed alphabetically. To complete the sample, additional digital libraries were found by browsing back issues of D-Lib magazine and performing a general Internet search. Each site was examined and considered for inclusion based on two criteria: (1) Did the project fit the study’s working definition of a digital library? and (2) Was contact information available to distribute the email survey?

Perhaps the most difficult task in designing and executing this study was operationalizing the term ‘digital library’. The literature abounds with competing definitions that focus on different aspects. Yet most definitions share commonalities. Saracevic (2000) surveyed definitions of digital libraries and summarized the common concepts. Digital libraries contain managed collections or resources that have been
selected, organized, licensed, secured and preserved. The resources have some type of representation or metadata and can be searched and retrieved. Users access digital libraries through interaction with interfaces via networks. Finally, digital libraries offer services in addition to collections. For the purpose of this investigation, sites were deemed digital libraries if they provided structured electronic collections of content, with metadata, and retrieval mechanisms, to a community of users in a distributed environment.

The second criterion for selection dealt with the practicality of identifying and contacting digital library project members. Two of the online registries provided a name, email, phone number, and address for the projects included in their databases. When this information was not available, the digital library was searched for a listing of team members. In some cases, the only way to contact a project member was to use a ‘comments and questions’ web form. When presented with more than one contact name, the project manager was chosen based on the assumption that he or she would be the most knowledgeable about any evaluations of the digital library. In most cases, however, there was only one name listed. In the event that the contact was not the appropriate person to ask, the survey requested that he or she forward the email to a staff person who could respond to the query.

Eighty-six digital libraries meeting this study’s inclusion criteria were identified from the Digital Library Federation’s online registry—a database of approximately 360 publicly accessible digital library projects sponsored by DLF members (http://www.diglib.org/pubs/techreps.htm). Each record in the registry provides
information about the digital library’s sponsoring organization, a brief description of the project, an email address for the project contact, and a link to the digital library.

Academic and Research Libraries maintains a database with 430 digital initiatives (http://www.arl.org/did/). Submissions can be any digital initiative in or involving libraries and are made on a voluntary basis. Seventy-two projects were identified for inclusion.

Eight more digital library sites were found on International Federation of Library Associations’ (IFLA) digital library resources web page. The site contains an annotated list of projects throughout North America and Europe.

Additional digital libraries were found by scanning all back issues of D-Lib magazine, 1995 through 2001, for articles featuring specific digital libraries. A search was entered in Google on the truncated term ‘digital librar*’. Browsing the first 75 entries in the retrieval set yielded ten more digital libraries.

The broad working definition for ‘digital library’ created a sample of projects with wide ranging characteristics. Due to the sampling sources, all digital libraries were available over the Internet and—with the exception of four sites—were either partially or totally accessible to the public. The collections consist of text, images, video and audio content on a variety of topics: art, religion, science, math, music, theater, history, cultural studies, language, politics and literature. Sites provided original content, non-copyrighted works, and third-party enhanced content. All digital libraries selected for the study contained some basic form of resource representation, such as title and creator information, and many provided more extensive metadata. All sites selected for inclusion had one or more mechanisms for retrieving resources. Searching capabilities ranged in
sophistication; some supported truncation and Boolean searching. Collections could also be browsed based on logical divisions in the material. The selected digital libraries serve a broad range of user communities. The targeted populations are explicitly stated in mission statements or implied by access restrictions based on membership criteria.

Seventy-two different institutions sponsor the digital libraries selected for inclusion. The majority, 43, are maintained by academic libraries. Seven digital libraries are federal sites and only six projects are developed by commercial interests. The remaining digital libraries are evenly divided between historical societies, public libraries, museums, and other public organizations.

Survey

The survey asked participants if their digital libraries had ever been evaluated. Since one of the objectives of the study was to determine how evaluation is being defined in practice, evaluation was broadly defined as any assessment of a system’s performance based on a specified objective. The survey also asked if those who had conducted some type of assessment would agree to a brief interview. The full-text of the survey is included in Appendix A.

Surveys were distributed and returned via email. 101 of the 200 projects responded to the survey. Seventeen of the surveys could not be delivered due to a problem with the email address and eighty-one had not been evaluated. Of the twelve projects (12% of the 101 returned surveys) that had conducted some type of assessment, eight agreed to an interview.
**Interview Participants**

Of eight project representatives who agreed to be interviewed about their digital library’s evaluation, two had evaluated two digital library projects each and discussed both assessments. One project contact replied that he had not evaluated the digital library about which the survey enquired, but would discuss the evaluation of a different digital project.

**Interview**

The interviews provided the main body of data for this study. Interviews were scheduled by email and conducted by phone. They consisted of thirteen questions (See Appendix B) designed to illicit a discussion about the motivations, methodologies, and outcomes of the evaluation.
Results and Discussion

Eleven case studies describe the projects and provide background information on their evaluations. They are followed by a discussion of the motivations, methodologies and outcomes of evaluation efforts. Finally, three themes that emerged from the interviews are discussed: adaptiveness, triangulation of data, and applying standard and well-defined measures.

Case Studies

Textbooks Digital Library

The Textbooks Digital Library began as a pilot project in 1997 when a university library with a large collection of historic textbooks offered online access to three digitized texts. Today, the site contains a collection of 33 textbooks, and a catalog of the entire print collection.

The site provides multiple search options including Boolean, proximity, bibliographic and simple searches, as well as browsing lists. After selecting a text, readers are presented with a short bibliographic record with author, title, publisher, physical extant, and hyperlinked subjects to other related texts. The metadata record is followed by a table of contents. Each page offers a set of viewing and navigation options. The user can set the page at 25, 50, 75, or 100% of the original page. Navigation

---

1 In order to protect the confidentiality of interview participants, pseudonyms are used for the digital library projects described.
features include the ability to go to the previous or next page, or jump to a specific page number, as well as a link to return to the table of contents.

Local History Digital Library

The Local History Digital Library is a joint venture between a university library and archives system, history department, and the local historical society. The collection includes over 300 full-text 19th and 20th century books, a 16 volume collection of maps with more than 650 plates, finding aids, a database of late nineteenth century census schedules, a chronology of local history, and a link to the historical society’s online catalog. The site is intended to benefit students, scholars, and historians who are studying the history of a city and its surrounding area.

The collection of texts is structured and searched in the same way as the Textbooks Digital Library. The maps can be searched or browsed by street and building name. After selecting a plate, the user is presented with a thumbnail of the image and permitted to set resolution and zoom settings. The census schedules can be searched on a number of fields including name, street, or the individual’s birthdate, occupation, and city of birth. The chronology can be searched by keyword and date or browsed using a clickable timeline to navigate through the chronology.

The textbooks and local history projects are both products of one university’s digital libraries initiative group. When the projects were first launched, there was no attempt to analyze logs or establish measures of use. A new department head joined the team and initiated log analyses and user surveys for both digital libraries to determine use and user population demographics.
**Music Digital Library**

Music Digital Library provides access to Latin music texts from the Middle Ages and Renaissance period. There are two options for accessing the collection. The browse option provides a listing of all texts by century with a brief metadata record and link to the text. The search option presents the user with a Boolean search interface and searching instructions. Search results contain information on the query (query time, number of retrieved items, total number of items), metadata, all instances of the search term in context, and a link to the full-text(s).

The Music Digital Library was evaluated by an outside agency interested in studying digital collections with the intent of creating a manual of best practices for digitizing cultural heritage collections. The evaluation was conducted as a series of interviews and work observations.

**Alcohol Studies Digital Library**

The Alcohol Studies Digital Library was developed by a university scholarly communication center and librarians with a subject speciality in Alcohol Studies. Over 55,000 research documents and educational materials pertaining to alcohol use are indexed. The search interface accepts keyword and controlled vocabulary Boolean queries. Users can search fields on physiological or social aspects, drug terms, author, date, resource format, or special populations.
Improving the retrieval mechanism was the primary objective for evaluating the Alcohol Studies Digital Library. The evaluator designed a tool to capture queries and result sets. The assessment focused on analysis of zero-hit searches.

Black History Digital Library

The Black History Digital Library is an award-winning site containing a sample of local black history materials from the state historical society. The extensive collection includes texts, letters, articles, and photographs. The search function permits Boolean searching and customizable display output. Or the collection can be browsed by subject or resource type. Searches return a bibliographic record including title, source, subject, call number and the repository from which it came with a link to the actual item.

This digital library was evaluated in an informal manner during the development process to get anecdotal feedback from colleagues. Because this digital library was funded by a grant, project deadlines prohibited a more formal evaluation. During the course of the interview, the project manager described another digital library, a State Scrapbook digital library, which has undergone a more formalized assessment.

State Scrapbook

The second project evaluated by this group was still under construction at the time of the interview. It is based on the Library of Congress American Memory site and brings together primary sources from a variety of institutions and sources to document the state history and culture from prehistoric times to the early twentieth century. Over three hundred universities, historical societies, public libraries, and other organizations
contributed source material. Using a scrapbook metaphor, the site is intended to be a central repository for archival materials documenting the state’s history and culture. The collection represents a wide range of media, including representations of manuscripts, organizational records, letters, diaries, photographs, biological specimens, art objects, furniture, clothing, and other objects of material culture. The collections are organized around five main topics: Citizenship, Economy, People, Culture, and Natural History.

The team conducted usability testing with 5 participants recruited from the local historical society reading room. The study focused on navigation, labeling, and metaphor usage. They also conducted a survey of participating institutions to ensure that the digital library met goals and expectations.

*Art Digital Library*

The Art Digital Library was the only museum-sponsored project to be examined. This site is dedicated to the photography of a twentieth century artist. Users can search a database of over 4,000 photographs, save specific images to a portfolio, or browse an archive of articles about the artist and his work. The Art Digital Library also contains educational activities and essays for children related to the life and times of the artist.

The Art Digital Library evaluation was conducted informally during the development process. Team members asked friends and colleagues to test the site, identify any problems, and comment on ‘look and feel’.
ImageBase

A university digital library research group found that managing a growing number of independently developed sites was expensive and difficult to manage. They restructured their digital initiatives into two templates: an image library template and a text library template. The digital library research group felt that this was an opportune time to evaluate the existing image digital library. In the fall of 2000, they invited a group of Information Science graduate students to conduct cognitive walkthroughs, heuristic evaluations, and interviews. The team followed up with an iterative cycle of prototyping and usability testing. ImageBase is the image template that emerged from that assessment and redesign.

ImageBase provides access to over 100,000 images from 29 public and restricted access collections. ImageBase has a wide range of search functions and features to manage result sets. The user can specify one of three display options, select four different images sizes, or zoom in and out by clicking on the image. A record displays below each image containing metadata on title, creator, date, description, subject, source, rights, and collection information.

Agricultural Digital Library

The Agricultural Digital Library project is a joint venture between a university agricultural library and federal agencies to provide public access to government agricultural reports and data sets. The site contains approximately 300 reports and datasets covering U.S. and international agriculture and related topics. Users can search for reports and data sets by keyword or browse by major subject headings. Each report
and dataset is annotated and offered in multiple formats. The digital library also offers a number of services including electronic reference, help desk support, and reports emailed on a regular basis to service subscribers.

The university library project members have undertaken a number of evaluations in an ongoing effort to identify trends in usage and user demographics. Transaction log analysis and electronic surveys have been their main methods of data collection.

*University Library Portal*

The University Library Portal serves as an online extension of one university’s libraries, archives, and scholarly communications department. It provides access to online journals, e-reserves, and theses and dissertations. The portal also houses a number of services and collections developed by the digital library team specifically for the university community. They maintain an archive of faculty publications and provide a website where researchers can post online surveys and collect the results.

The digital library team uses this website to post their own surveys to determine if their services and products meet users’ needs. They also collect anecdotal information the site in user education training sessions.

*Video Digital Library*

The Video Digital Library began in 1994 as an experiment in automatic indexing, visualization, and retrieval for video and audio materials. Currently the collection contains 1,500 hours of news footage and documentaries. The library offers full-content searching on audio and closed captions with some image matching for faces.
The Video Digital Library was the only research-oriented project addressed in this study. As an experimental system, evaluations have been conducted on a regular basis as a part of the research process. The interview focused on a study of visualization techniques used to represent entire video segments with a single thumbnail image.

Motivations

The digital library projects examined in this study conducted assessments for many of the same reasons. For several of the digital libraries, the evaluations occurred as a part of the development cycle. The three projects that performed informal evaluations were all motivated by the need for timely, objective feedback on a site that was close to release. They all stated that they were so closely involved in the project that they wanted non-biased opinions from outsiders. The ImageBase team performed a formal, summative evaluation to assess the strengths and weaknesses of the existing image digital library before designing a new system. Later in the development cycle, users were brought in to test prototypes.

Another common reason for evaluation was to identify users and their information needs. The evaluators of the Agricultural Digital Library—which is an extension of a physical library—wanted to know how delivering the same content in an electronic environment would affect use and user demographics. The project leader explained,

The reason that we would do any evaluation would be to see if the system was being used, to discover who our users are. We have a good sense of who in a traditional sense would come to an agricultural library... [W]e began to see that we would no longer be a library that somebody had to drive up to, but someone could connect to from anywhere from Albania to Zimbabwe. We wanted to know, is this really happening?
The Historic Textbooks project was surprised to discover that, in addition to the expected user populations—historians and scholars—an overwhelming number of visitors were homeschoolers.

Gaining a better understanding of users’ information needs was often cited as a related motivating factor. The Local History Digital Library considered expanding their collections to different resource types and wanted to know which resources the users wanted most. The Agricultural Digital Library team wanted to know what data formats were most useful. The State Scrapbook team members were not only concerned with the needs of users; they also wanted to make sure that they met the expectations of all the partner institutions that contributed source material.

Projects were also evaluated as a form of documentation. The Black History Digital Library, Historic Textbook Digital Library, and Local History Digital Library all conducted summative evaluations, because new management wanted to assess the current state of the projects. The Music Digital Library underwent an assessment as part of a larger effort to establish a ‘best practices’ manual.

Several projects mentioned that data collected from evaluations was useful in reporting progress to financial backers and stimulating funding. This purpose was almost always aligned with usage statistics. The Historic Textbook project received additional funding to digitize 100 texts when web logs indicated that the number of site hits exceeded all expectations. The Agricultural Digital Library employed usage statistics to justify archiving older reports when their administration wanted to limit the scope to current documents only.
Objectives

The evaluation objectives were unique to each project, but fall into three broad categories: more effective retrieval mechanisms, enhanced interfaces, and improved collection and services.

The Alcohol Studies Digital Library evaluator focused most of his attention on improving the retrieval mechanism. He designed a facility to capture search terms and retrieval sets. After collecting this data over a six-month period, he determined that 32% of the searches returned zero hits. The evaluator believed that improved searching instructions and the use of a controlled vocabulary could lower the percentage of zero-hit searches.

The State Scrapbook conducted usability testing to assess aspects of the interface. They wanted to know if users could navigate the site, understand terminology and labeling, and grasp the scrapbook metaphor.

The Video Digital Library and ImageBase assessments addressed both retrieval and interface design. The Video Digital Library team designed an automatic indexing mechanism that selected a single thumbnail image to represent an entire video clip and developed a visualization technique to present these thumbnails in a result set. They conducted a usability test to determine how long it took for their participants to retrieve specified clips and how satisfied they were with the interface. The ImageBase evaluation studied how users went about retrieval tasks in an image database and how aspects of the interface could be utilized to make this task easier.

The University Portal, Agricultural Digital Library, Local History Digital Library, and Historic Textbook project all aimed to enhance some aspect of their collection or
services. The University Portal surveyed users to assess user satisfaction of the electronic theses and dissertations database and e-reserves. The Agricultural Digital Library sought to understand how users were using the collections in order to provide data in more useful formats. Finally, the Local History and Historic Textbooks Digital Libraries focused their efforts on identifying which areas of the collection that should be expanded.

Methodologies

The digital libraries selected for this study employed transaction log analyses, surveys, and usability testing to evaluate their sites. Three projects used a combination of surveys and log analyses and one site conducted usability testing and surveys. Two digital library projects conducted usability testing, one focused on log analysis and another on surveys.

Surveys

All projects administered and collected their surveys through a web interface. The Agricultural Digital Library also emailed the survey and provided multiple options for returning the completed survey—email, web interface, post, or fax. The evaluator explained that the multiple options made it more difficult to collect and analyze the data. In the next survey, the instrument will only be administered through a web interface. In all but one case, there was no attempt to selectively sample respondents. The Agricultural Digital Library selected the survey sample from their email report service to target repeat users.
The number and content of survey questions varied by project, but they all asked basic questions: Who are you?, What information are you looking for?, What additional resources or services would you like us to provide?, and, What is your level of satisfaction with the digital library?

There may be several explanations why surveys were the most commonly used data collection method. They are inexpensive, easy to implement, analyze, and require very little effort to coordinate. They also provide more detailed information about user demographics and information needs—a key consideration for almost all the digital libraries studied here—than any other methodology.

However surveys cannot be effective evaluation instruments if the questions are not designed carefully. Evaluators from the Agricultural Digital Library and the Local History Digital Library described experiences that underscore this fact. The Agricultural Digital Library team members spent several months designing, evaluating, and refining their survey questions. They employed the services of a communications graduate student who had significant experience in survey design. As a result they collected meaningful data. The project coordinator expressed his satisfaction with that aspect of the evaluation: “A well designed survey gives you information that you can go back to again and again. I think that we did it right and I think it will probably get a whole new life of its own because the next survey will be related to what we learned in the first one.”

The Local History Digital Library evaluator felt that several of the survey questions were ambiguous or poorly conceived. As a result he did not find them as useful. When asked if he felt the assessment had been a success he gave a qualified response. He said that for
his purposes it was successful, but “we could probably do a better job with a future survey if I worked with a professional”.

Transaction Logs

Four digital libraries reported using transaction logs in their assessments. Two of the projects used the standard web log generated by their Apache server and two projects wrote their own applications. The Historic Textbook and Local History Digital Libraries relied on standard Apache web logs to provide number of hits as a rough measure for usage. The evaluator stated that he used these statistics to report the projects’ progress to administration more than as an assessment tool. “I think that the transaction logs . . . have to be interpreted very carefully because even though [it reports] this number of unique IP addresses or that number of hits, the numbers don’t always mean what you think they mean.”

The Agricultural Digital Library team wrote their own log application, which weeded out what appeared to be anomalous lines and collected some of the basic usage statistics. It was also designed to record patterns in information clustering. The log looked for correlations in use of older reports with the most current reports.

The Alcohol Studies Digital Library evaluator used log analysis as the main source of data for an assessment of the retrieval mechanism. He created an analysis tool to capture users’ search terms and the number of items in the result set. Focusing on searches that retrieved zero hits, he analyzed the semantics and syntax to discover ways to improve the search feature. He pointed out that logs alone cannot differentiate between search sessions. Since the site has no login mechanism, he had to designate a
search session as searches performed by a single IP address within a twenty-minute interval.

The teams that used web logs characterized them as convenient tools that provide ready data. However, they are not easily interpreted and provide only a rough gauge of usage. User demographics, for example, cannot be gleaned from log analysis.

Usability Testing

Three projects employed usability testing in their evaluations. This method of data collection was cited as the most labor and resource intensive, but it provided important information that could not be gained from other methodologies. All the projects that conducted usability testing named an improved interface as their main objective for the assessment.

The ImageBase team took advantage of its close proximity to an Information Science school and asked a group of students to evaluate the site for a term project. The group did a heuristic evaluation, cognitive walkthroughs, user interviews, and some small scale user testing. Based on the students’ report, the developer created model mockups for the site. Using this prototype, the team asked fifteen participants to complete a series of tasks that they felt were representative of typical information needs. The tasks included searching for specific items in the collection, using a personal data collection space called a portfolio, and finding information in the metadata record. Feedback from a usability session was integrated into design specifications, an improved prototype was developed, and the process started over again.
The Video Digital Library employed usability testing to analyze their retrieval mechanism and interface. College students were recruited to complete a set of retrieval tasks. The team measured how long it took to retrieve specified images and how satisfied the students were with the interface. The evaluator commented that one of the biggest obstacles in designing this study was to come up with non-biased questions. Their questions lean toward the visual, automatically biasing any interface that presented textual result sets. To overcome this problem, the team adopted questions from a program that teaches English as a second language from CNN news footage.

The State Scrapbook project conducted usability testing to assess the site’s navigation, labels, and the scrapbook metaphor. They set up a testing center in the reading room of the local historical society with the intention of testing “the person off the street”. While the usability testing did yield useful results, the evaluators found they had trouble recruiting participants on the spot. User testing is a more laborious process, but provides valuable “real-world” interactions between the user and the system.

Additional Methodological Issues

The digital library teams learned a great deal about the evaluation process as they design and execute assessments. If this information is published and shared with other projects, then evaluators can learn from others’ successes and mistakes, benefiting the digital library community as a whole. Are digital libraries reporting the results of their evaluations? Two of the more informal evaluations were not reported at all, in keeping with the purpose of collecting quick feedback. Two projects published their reports online. One group made the report available on the project’s website and the other
distributed an article in the university’s online newspaper. One group published their results in a conference proceeding. One investigator intends to submit a paper on the evaluation effort to a journal.

The teams reported that logistical problems and difficulty in designing effective data collection tools were the biggest obstacles in obtaining useful data. Two groups cited problems recruiting volunteers. The State Scrapbook team solicited volunteers for a usability study from patrons walking in to the library. They found it difficult to get people to participate and had trouble gauging the level of technical comfort among those patrons who did participate. The ImageBase evaluator felt that recruiting volunteers without having the budget to offer financial incentives and difficulties around scheduling 25 volunteers and digital library team members was problematic. Two evaluation efforts had difficulty designing surveys that solicited the information that they wanted from the evaluation. One project leader reported that there were questions they would have liked to ask but could not because they were not sure how to frame the question to get a meaningful and consistent response. The Local History Digital Library team reported that the survey was not specific enough to produce useful data; the digital library encompassed a large collection of varied materials and the questions posed in the survey asked broad questions. After the first survey results were collected, another survey was posted with different components for each part of the digital collection.
**Evaluation Characteristics**

Digital library researchers who have conducted long-term evaluations describe three characteristics of a beneficial assessment. Evaluations of a new and evolving technology are exploratory in nature. Evolution is a process over time; any investigation that attempts to understand that evolution should be a long term undertaking. Furthermore, evaluators should incorporate multiple measures into their assessments. New insights can be gleaned from unexpected relationships between data sets, i.e., triangulation. In turn, these new insights influence the course of the evaluation, requiring a flexible and adaptive design (Marchionini, 2000). While this is an excellent paradigm to strive for, this study suggests that it may not be an attainable model for many practicing digital libraries.

*A Long-term Outlook*

Ideally, evaluation would be an integral part of a digital library’s life cycle. In reality, few projects have the funding or staffing to devote to such an effort. Most of the digital library teams studied here were responsible for multiple digital libraries. In some cases, they moved from the launch of one site directly into the development of a new project. The Agricultural Digital Library and the Video Digital Library were the only projects with a history of ongoing assessment cycles.
Adaptiveness

The second important evaluation characteristic—adapting the assessment as new information emerges or the technology changes—cannot occur unless the project can be studied over time. However, several of the digital library teams applied what they had learned from the evaluation of one digital library to their next evaluation effort. In this respect, the evaluations under investigation displayed characteristics of adaptiveness.

Triangulation

The majority of the evaluations described in this study relied on a single methodology. Three projects incorporated surveys and logs into their evaluation and one project utilized surveys and usability testing. Two of the four projects actually triangulated data. The Local History Digital Library compared users’ reported information needs from surveys with document retrieval statistics gleaned from the logs to identified user information needs. The Agricultural Digital Library evaluator used log analysis to find retrieval patterns in document clusters and surveys to determine user demographics and information needs. The team will use the results from this investigation to guide the next development cycle.
Conclusion

While awareness is growing, evaluation is still not an integral part of digital library development. Only twelve of the two hundred projects surveyed have been evaluated in some way. The eight projects interviewed conducted their evaluations to identify their user population or provide evidence of progress to administration. The nature of the assessments ranged from informal feedback from friends and colleagues to formalized usability tests conducted in a laboratory. The interviews indicate that most evaluations were a short-term activity closely tied to development. Most evaluations were too brief to exhibit characteristics of adaptability, but the evaluators did incorporate what they had learned from the evaluation of one digital library project into the assessment of others. Finally, there is little evidence that digital library evaluations incorporate triangulation in data collection and analysis. Only four of the ten projects interviewed used more than one method of data collection. However two of these projects utilized some form of triangulation.

The survey responses indicated a strong interest in evaluation. Several individuals from projects that had not previously conducted an assessment indicated that they would be interested in undertaking one, while others requested a reading list on the topic. The digital library projects interviewed for this study are leading the way for others.
What steps can be taken to capitalize on this interest and support the work of digital library teams who are already conducting evaluations? Sponsoring institutions and financial backers can show their support by establishing regular funds for on-going assessment activities. The evaluators interviewed in this study already have many job responsibilities and took on the additional duties associated with evaluation out of personal interest and professional commitment. Other digital library teams may not be able to take on such tasks without additional staffing and resources.

Digital library projects that have conducted evaluations can share their experience and encourage others to begin assessments by publishing their results. Many of the evaluators interviewed did not feel that they had the time to write a formal paper or felt that the methodology was not rigorous enough to be published in a peer-reviewed journal. These individuals published their findings on their websites or in internal publications.

Digital library associations and working groups could make it easier for evaluators to publish their findings by maintaining a web database of evaluation case studies. Digital library teams could quickly and easily submit their reports on a voluntary basis. Other digital library projects interested in undertaking assessments would not have to ‘reinvent the wheel’; they could build on the experiences of others. A centralized repository would also make it easier for researchers in the digital library community to identify trends and establish standards for evaluation.

Digital libraries are a rapidly evolving technology. Some have envisioned digital libraries evolving toward a common structure in the same way that all physical libraries have the same organization (Saracevic, 2000). In order to guide digital libraries toward
this goal, the digital library community needs to systematically and collectively reflect on their work.
References


Appendix A
Email Survey with Consent Information

Dear [Participant],

I am a student at the School of Information and Library Science—UNC-CH, conducting research on digital library evaluation for my Masters paper (see description below).

During the life of [Name of Digital Library], has the system ever been evaluated? Evaluation is any assessment of the system’s performance based on a specified objective. Examples include-- but are not limited to-- usability studies, transaction log analyses, or surveys.

If you have done evaluation, would you agree to a follow up phone interview that will take no more than 30 minutes? I will ask questions about why and how the system was evaluated and how the results were (are being) incorporated into development activities. Even if you have not done evaluation or do not wish to be interviewed, I would greatly appreciate a reply.

If you cannot address these questions, would you please forward this request to the [Name of Digital Library] project coordinator or other appropriate staff?

Thank you for your time and consideration.

Sincerely,

Holley Long

Description of Research Project:

Assessing the Current Role of Evaluation in Digital Library Development

This study will provide a snapshot of current digital library evaluation activities. Evaluation is any assessment of the system’s performance based on a specified objective. Examples include-- but are not limited to-- usability studies, transaction log analyses, or surveys. What percentage of digital library projects are assessing their sites? How is evaluation conceived and carried out? How are the results being used in development? Is evaluation perceived as a useful activity?
Approximately 200 digital library project managers are being surveyed by email to determine what percentage of these projects have undergone evaluation. Up to twenty of the respondents who agree to answer a few follow-up questions will be randomly selected for an interview. If you chose to participate in the survey or interview, your responses will be completely anonymous. While this email and your reply message are not encrypted, every effort will be made to ensure your privacy. Once the data have been collected, all emails, interview tapes and transcripts that identify you or your digital library will be destroyed. As a benefit to those who participate, I will email a summary of my findings to any respondent who wishes to receive it.

If you have questions about this research, please contact: Holley Long (principle investigator) at longh@ils.unc.edu or 919-968-2179; Barbara Wildemuth (faculty advisor) at wildemuth@ils.unc.edu or 919-962-8072.

If you have concerns regarding your rights as a research participant, please contact:

    Academic Affairs Institutional Review Board:
    Barbara Goldman, Chair
    CB# 4100, 201 Bynum Hall
    University of North Carolina at Chapel Hill
    Chapel Hill, NC 27599-4100
    919-962-7761 or email: aa-irb@unc.edu
Appendix B
Interview Questions

What motivated you to evaluate the digital library?

What were your objectives for evaluation?

What aspects of the system were evaluated?

What evaluation methodologies were used? These might include usability testing, log analysis benchmarking, or any other type of assessment.

If multiple methodologies were used, did you find one to be more or less informative? Why?

What performance measures were used?

What were some of the biggest obstacles in obtaining useful data from the evaluation?

During the time frame you conducted evaluation, what percentage of your time was devoted to this activity?

How did the evaluation findings influence development activities?

How and to whom were the results reported?

Given your objectives for evaluating the digital library, do you feel that the evaluation was successful? Why or why not?