Paper submitted to the *Annual Meeting 2000 of the American Society for Information Science*

Challenges for Digital Library Evaluation

Tefko Saracevic, Ph.D.
Lisa Covi, Ph.D.

School of Communication, Information, and Library Studies

Rutgers University

4 Huntington Street, New Brunswick, NJ 08901
tefko,covi@scils.rutgers.edu

Abstract

While there were many efforts in research and practice of digital libraries, evaluation was not a conspicuous activity. It is well recognized that digital library evaluation is a complex and difficult undertaking. We enumerate the challenges facing digital library evaluation and suggest a conceptual framework for evaluation, A review of evaluation efforts in research and practice concentrates on derivation of criteria used in evaluation. Essential requirements for evaluation are stated. Discussed are constructs, context and criteria of digital libraries: What should we evaluate? For what purpose do we evaluate? Who should evaluate? At what level do we evaluate? Upon what criteria do we evaluate? In addition we include suggestions for adaptation of criteria from related activities. The paper is considered as a part of the evolution of concepts for digital library evaluation.

Introduction

Digital libraries have a short, yet turbulent and explosive history. A number of early visionaries, such as Licklider (1965), had a notion of libraries in the future being highly innovative and different in structure, processing, and access through heavy applications of technology. But, besides visionary and futuristic discussions and highly scattered research and developmental experimentation, nothing much happened in the next two decades. By the end of the 1980's digital libraries (under various names) were barely a part of the landscape of librarianship, information science, or computer science. A decade later, by the end of the 1990's, research, practical developments, and general interest in digital libraries exploded globally. What a phenomenal decade for work on digital libraries! The accelerated growth of numerous and highly varied efforts related to digital libraries continues unabated in the 2000's. While the exciting history has yet to be written, Borgman's (1999) discussion of competing visions for digital libraries is a good beginning for understanding the forces and players involved. These competing visions and associated definitions come from several communities that are involved in digital library work. We are concentrating here on two communities: research and practice, as reviewed below. They work and proceed mostly independently of each other.

Large resources and efforts have been expended on digital library research and practice. There are many efforts, projects, and implementations, not only in the U.S. but in many other countries, and on international levels as well. More are under way. A great many things are being done and explored, but evaluation is conspicuous by its absence in the vast majority of published work on digital libraries, be it research or practice. So far, evaluation has not kept pace with efforts in digital libraries, or with 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. At this stage of digital library evolution, evaluation in any formal sense (as opposed to anecdotal) is being more or less bypassed. True, evaluation has been talked about and implemented in a few instances (as reviewed below), but these are exceptions rather then a rule. Why is that?

We can only speculate as to the answer. Perhaps it is too early in the evolution of digital libraries to attempt evaluation in any formal way. Informal and anecdotal ways suffice. Maybe evaluation is taken to be sufficient on a very basic technical level - the fact that an electronic collection is searchable, accessible and used is evaluation in itself. From a cynical perspective, we might suggest that the interest in evaluation is suppressed - who want to know about or demonstrate the actual performance? Perhaps the performance is evident in the use and popularity? On the other hand, perhaps in the pressure of the rapid pace of evolution, the rush to do something and then to rush to something next, does not leave time and room for evaluation. And maybe evaluation of digital libraries is so complex that even when desired, it cannot be accomplished with what we presently know about evaluation. In other words, we might conclude that the conceptual state-of-the-art of digital library evaluation is not sufficiently developed to start with.

While all these speculations may be true to some extent, we believe that the last, the one about the underdeveloped conceptual nature of evaluation, is actually true. Evaluation of digital libraries is a complex undertaking and thus, it is a conceptual and pragmatic challenge. In this paper we are addressing various conceptual and theoretical questions about evaluation of digital libraries and proposing concepts and approaches that we believe to be appropriate toward their evaluation. We consider this paper as a part of the evolution of the concepts for digital library evaluation.

Digital library communities

While there are numerous communities interested in digital libraries we are, as mentioned, we are concentrating here on the research and practice communities as being most closely evaluation bound. Each has a differing conception and definition, affecting the conceptual nature of evaluation. This translates into specific questions:*What is a digital library? What is there to evaluate? What are the criteria? How to apply them in evaluation? Why to evaluate digital libraries in the fist place?*

Research community

The research community, whose most vocal members reside in computer science, concentrates on developmental research and some experimentation dealing with technology applications in a variety of areas and media, for various communities, and on enabling technologies and networks as an infrastructure for digital libraries. While there is a notion that the research will result in practical applications and in actual digital libraries, the goal is not connected to actual operations, but to research - this is an important point to consider because it impinges on evaluation. In the U.S., digital library research is guided and even defined through the projects supported by Digital Library Initiatives (DLI). The Initiatives are funded by a consortium of government agencies under the leadership of the National Science Foundation (NSF). DLI-1 (1994-1998), funded by three agencies, involved six large projects. DLI2 (1999-2003), funded by eight agencies, involves close to 60 large and small projects. There are also large research digital library initiatives in the U.K., Germany, the European Union, Japan, and elsewhere. In this paper we concentrate only on the efforts in the U.S., while recognizing the existence of many other efforts in many other countries.

Digital Library Initiatives did not define 'digital library.' In order to incorporate a wide range of possible approaches and domains, the concept is treated broadly and vaguely. Thus, the projects, particularly in DLI2, cover a wide range of topics, stretching the possible meaning of 'digital library' to and even beyond the limit of what can be considered as being 'digital' and at the same time recognizable as any kind of a 'library' or a part thereof. This is perfectly acceptable for research - frontiers need to be stretched. But at the same time, it makes evaluation not exactly a possibility, to start with. It is not surprising then that evaluation is hardly a significant part of DLI efforts.

While formal evaluation was not a big part of DLI-1, three interesting approaches merged. The most notable formal evaluation was done within the Alexandria Digital Library Project (ADL) at the University of California, Santa Barbara (Hill et al, 2000). The approach involved a series of user studies, involving different user communities and concentrating on different design features as related to their usability and functionality. Some of the results were fed back to

improvements of design, "influenc[ing] the Project's implementation goals and priorities." The results served as a base for specifying a " partial list of requirements for new ADL interfaces that came from user evaluation studies." User logs were also studies as a part of the evaluation. The evaluation concentrated on users and their interactions through the interface, with *usability* and *functionality* as the main criteria. The usability studies have become one of the more popular and implementable ways to approach digital library evaluation (e.g. Buttenfield, 1999). But, usability is only one of the possible and needed criteria and approaches.

In the DLI-1 project at University of California at Berkeley as part of evaluation a series of interview with intended users were conducted (Schiff, Van House & Butler, 1997). They focused on situated actions, defined as "[action] performed by specific individuals in specific socio-cultural context using tools and technologies for a specific purpose." A sociological theory about the relationship between individual agency and fields of behavioral orientation by Pierre Bourdieu (1990) was used as a framework. They concluded, "investigating the social setting for which a DL is intended provides us with a rich understanding of the people involved, their relative interest and abilities to act, their opportunities and constraints, and their goals." The criteria for the study of users are *social environment* and *user actions*. However, it is not clear whether Bourdieu's theory of "habitas" can be immediately applied and used to test digital libraries.

In the DLI-1 project at the University of Illinois academic researchers studied how readers use scientific journal articles in both print and digital environments - how they "mobilize the work ... as they identify, retrieve, read and use material in articles of interest"(Bishop, 1999). The criteria were *work* and *use of retrieved materials by users*.

The last two, and similar studies of user behavior related to digital libraries or to information in general, provide useful information, as pointed out by Bishop "[with] implication for digital library system design and user education." But, they are really not directly devoted to systematic evaluation. This raises the larger point: User studies, while useful for understanding how people use systems, by themselves are <u>not</u> evaluation, even though they may have evaluative implications and they provide important criteria that can be used in evaluation.

Practice community

The practice community, whose majority is residing in operational libraries, concentrates on building operational digital libraries, their maintenance and operations, and providing services to users. The approach is eminently practical, with relatively little research involved. As a result, hundreds, if not thousands of digital libraries have emerged worldwide, with more becoming operational every day. The efforts are diverse. Many approaches are being used. Many types of collections and media are included and processed in many different ways. Many are located in libraries, creating a hybrid library (combination of a traditional and digital library), others are not bound to libraries at all. The Library of Congress on its web pages provides an impressive set of links to various digital libraries (starting with www.loc.gov), and so does the journal D-Lib Magazine (www.dlib.org).

One of the earliest and longest lasting evaluation of practical digital libraries is the evaluation of the Perseus Project - a corpus of multimedia materials and tools related to the ancient Greek world, (www.perseus.tufts.edu). The mission of Perseus is to provide improved access to primary source materials related to needs of students and faculty to to better learn and understand culture. The evaluation addressed a set of questions related to learning, teaching, scholarly research in the humanities, and electronic publishing (Marchionini & Crane, 1994). Four evaluation criteria were were identified: *learning, teaching, system (performance, interface, electronic publishing), and content (scope, accuracy)*. The evaluation provided a number of results that were summarized in four categories: amplification and augmentation of learning, physical infrastructure, conceptual infrastructure, and systemic change to the field. This is still a model evaluation project for digital libraries.

PEAK, standing for "Pricing Electronic Access to Knowledge," is one of the more interesting projects that involves both observation of use and evaluation of a variety of aspects, particularly including economic factors (Bonn et al. 1999;

Mackie-Mason et al. 1999). It is unique in that involves a publisher of electronic journals, Elsevier Science, and about a dozen libraries. (A project TULIP also done by Elsevier and a number of universities preceded PEAK). Criteria for evaluation included *access* (different types of access to journals was offered to different groups), *pricing* (different models), and *revenues and costs*. This project extended evaluation criteria and measures to economic or efficiency evaluation.

The Museum Educational Site Licensing (MESL) Project was a collaboration of seven collecting institutions and seven universities, defining the terms and conditions for the educational use of digitized museum images and related information. The MESL implementation at Cornell, as a separate digital library, has been conducted under the auspices of the University's Digital Access Coalition. A report describes the implementations and some evaluation (Cornell, 1999). The approach is impressionistic - a lot of questions have been asked of users, designers, developers, and operators to obtain evaluative impressions and assessments. Criteria in questions include: *functionality - browsing, searching, difficulty, usage, experiences, training needs, integration into other campus services, preparation of source materials for inclusion, fields indexed, server performance, system security and authentication, ongoing support needed or desired, technical development, physical infrastructure, costs, time, skills.* The evaluation was not formal, but it is interesting if for nothing else then for the breadth of criteria included.

Since 1995, the Human-Computer Interaction Group at Cornell University has conducted research or evaluation studies of a number of prototype efforts to build digital collections in museums and libraries (Jones, Gay & Rieger, 1999). In that paper they summarized five studies. The criteria used revolve around: ""backstage" concerns" or *representation, legal issues* ("e.g., metadata, copyright and intellectual property issues"), *collection maintenance and access* ("e.g., decisions regarding collection scope and the maintenance of a consistent quality and fidelity of digital records"), and *usability* ("e.g., user skill levels and expectations, and the use of collections in formal and informal educational settings"). The methods used in these evaluations are not clear, to what degree were they formal or informal. But a number of conclusions were drawn, among them:

> Effective digital collections are complex sociotechnical systems: An effective collection requires consistent and simultaneous attention to a variety of social, organizational, administrative, and technical concerns.

A number of other authors came to the same conclusion illustrating a model of digital libraries that involves a wide range of levels, as suggested below.

In this review we included only representative efforts, primarily to illustrate criteria used. But, we could not find many other evaluations in either research or practice. This illustrates our point that there is a dearth of evaluation efforts. In addition, practical evaluations may be motivated by the desire to promote use of constructed resources and anticipated outcomes of use.

Needed and lacking for digital library evaluation

The general questions in any and all evaluations are:

Why evaluate? What to evaluate? How to evaluate?

There are many approaches to evaluation and to answering these questions. We fully recognize the appropriateness of different approaches for different evaluation goals and audiences. For instance, the ethnographic approach is highly appropriate for gaining a broad understanding of the role and effects of a practice or a construct in a wider social or organizational framework. The sociological approach is appropriate to illuminating the social forces and effects.

However, here we concentrate on the systems approach only, as the most widely practiced or suggested approach for evaluation of all kinds of information systems, including digital libraries. We start with the basic assumption of all systems approaches to evaluation: evaluation deals with some aspect of performance. Thus, the general *why* of evaluation deals with performance to start with and goes on from there to define more specific goals and choices, as discussed under context of evaluation below.

A system can be considered a set of elements in interaction. A human-made system, such as a digital library, has an added aspect: it has certain objective(s). The elements, or the components, interact to perform certain functions, or processes, to achieve given objectives. Furthermore, any system (digital libraries included) exists in an environment or environments (which can also be thought of as systems), and interacts with its environments. It is difficult and even arbitrary to set the boundaries of a system. In evaluation of digital libraries, as in evaluation of any system or process, these difficult questions arise that clearly affect the results: *Where does a digital library under evaluation begin? Where does it end? What are the boundaries? What to include? What to exclude?* This sets the questions of determining the construct of digital libraries, as discussed below.

In this context, by evaluation we mean an appraisal of the performance or functioning of a system or part thereof, in relation to some objective(s). The performance can be evaluated as to

- ← effectiveness: *How well does a system (or any of its parts) perform that for which it was designed?* Or as to
- ← efficiency: *At what cost?* (Costs could be financial, time or effort). Or as to
- ← a combination of cost-effectiveness.

An evaluation has to specify which of these will be evaluated. From now on, we will mostly discuss evaluation of effectiveness, with a realization that at any evaluation efficiency, and cost-effectiveness can be involved as well. This sets the questions of the criteria of evaluation for digital libraries, as discussed below

As in all systems, objectives occur in hierarchies, and there may be several hierarchies representing different levels - sometimes even in conflict. While the objectives may be explicitly stated or implicitly derived or assumed, they have to be reflected in an evaluation. Evaluation is not one fixed thing. For the same system, evaluation can be done on different levels, in relation to different choices of objectives, using a variety of methods, and it can be oriented toward different goals and audiences.

To be considered as an evaluation, any evaluation has to meet certain requirements. It must involve selections and decisions related to:

1. Construct_for evaluation. *What to evaluate? What is actually meant by a digital library? What is encompassed? What elements (components, parts, processes…) to involve in evaluation?*
2. Context of evaluation - selection of a goal, framework, viewpoint or level(s) of evaluation. *What is the level of evaluation? What is critical for a selected level? Ultimately: What objective(s) to select for that level?*
3. Criteria reflecting performance as related to selected objectives. *What parameters of performance to concentrate on? What dimension or characteristic to evaluate?*
4. Measures *reflecting selected criteria to record the performance. What specific measure(s) to use for a given criterion?*
5. Methodology for doing evaluation. *What measuring instruments to use? What samples? What procedures to use for data collection? For data analysis?*

A clear specification on each of these is a requirement for any evaluation of digital libraries. Unfortunately, it is not as yet entirely clear what is to be specified in each of these five elements. No agreement exists not only on criteria, measures, and methodologies for digital library evaluation, but even on the 'big' picture, the construct and context of evaluation. The evaluation of digital libraries is still in a formative stage. Concepts have to be clarified first. This is the fundamental challenge for digital library evaluation.

A clarification is needed as to what does not fall in the realm of evaluation, even though it could be related to evaluation. By itself measurement or specification of metrics for digital libraries is not evaluation - it is a quantitative or qualitative characterization. Observation by itself, such as observing user behavior in the use of a digital library, is not

evaluation. Assessing user needs by itself is not evaluation, and neither is relating those needs to design. However, these can be linked to evaluation if and only if they are connected to some specified performance by including all the five requirements enumerated above.

A related view of evaluation is expressed by Marchionini, Plaisant & Komlodi (in press):

> Evaluation of a digital library may serve many purposes ranging from understanding basic phenomena (e.g., human information-seeking behavior) to assessing the effectiveness of a specific design to insuring sufficient return on investment. Human-centered evaluation serves many stakeholders ranging from specific users and librarians to various groups to society in general. Additionally, evaluation may target different goals ranging from increased learning and improved research to improved dissemination to bottom line profits. Each of the evaluation goals may also have a set of measures and data collection methods. Finally, the evaluation must have a temporal component that can range from very short terms to generations.

Construct: What is a digital library?

What is there to evaluate? A simplistic answer is that whatever is called by those involved a 'digital library' is a digital library, thus a construct candidate for evaluation. (This is derived from the philosophy whose metaphor is "Physics (or whatever field) is what a physicist does"). This is a pragmatic approach that has been applied to modeling a construct of a digital library, and to some extends it works. But a more formal approach to defining or modeling the construct is needed in order to develop generalizations to and from evaluations.

Because digital libraries are related to physical libraries, and may perform a number of similar functions, but in relation to digital and distributed collection, the modeling and evaluation of digital libraries may to some extend parallel that of physical libraries - at least initially. But, digital libraries are also sufficiently different, and in some functions, as for example in distribution and access, completely different than physical libraries. Thus, digital libraries also require additional and new approaches to modeling of their constructs and to evaluation. Also, a digital library is much, much more than a collection of digitized texts, and other objects. The challenge to digital library evaluation is developing and applying these new modeling and evaluation concepts and approaches.

As mentioned, in the research community 'digital library' has not been defined. The closest to the definition applicable to the approaches taken by the research community is the one given by Lesk (1997) in the first textbook on the topic:

> digital libraries are *organized collections* of digital information. They combine the *structure and gathering of information*, which libraries and archives have always done, with the *digital representation* that computers have made possible. (Emphasis added).

The emphasized elements in the definition represent constructs that could and should enter in evaluation, answering the question at the start of this section. The question should be raised: is this enough? We do not think so.

Borgman (1999) provides a more complex definition (including an extensive discussion) of digital libraries, a definition that may be considered as a bridge between the research community definition above and practical community definition below:

1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. … they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium …The content of digital libraries includes data, [and] metadata …
2. Digital libraries are constructed, collected, and organized, by (and for) a community of users, and their functional capabilities support the information needs and uses of that community.

In this definition, the elements in the construct subject or candidates for evaluation are:

- ← electronic resources, digital data in any medium,
- ← technical capabilities for creating, searching, and using information,

- information retrieval,
- metadata
- community of users; their information needs and uses.

In the US, the Digital Libraries Federation (DLF) is an organization of research libraries and various national institutions formed in 1995. The stated goal of DLF is "to establish the conditions necessary for the creation, maintenance, expansion, and preservation of a distributed collection of digital materials accessible to scholars and the wider public." The organization represents the practical community. After considerable work, DFL agreed on a "working definition of digital library," representing definition of the practice community:

> Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. (DFL, 1999)

This definition and conception, is quite different from the one provided by Lesk (1997), and even by Borgman (1999). Here, the emphasis is on organizational or institutional setting for the collection of digital works and aspects related to its functioning in the larger context of service, which specifically involves these elements in the construct subject or candidates for evaluation:

- *professional staff,*
- *collection of digital works*
- *selection, structure, and access,*
- *interpretation and distribution,*
- *preservation, and*
- *use and economic availability for a defined community*

Let us attempt integration. In a general way, the constructs or elements in digital libraries, candidates for evaluation are:

- Digital collections, resources
  - selection, gathering, holdings, media
  - distribution, connections, links
  - organization, structure, storage
  - interpretation, representation, metadata
- Preservation, persistence
- Access
  - intellectual
  - physical
  - distribution
  - interfaces, interaction
  - search, retrieval
- Services
  - availability
  - range of available services e.g. dissemination, delivery
  - assistance, referral
- Use, users, communities
- Security, privacy, policies, legal aspects, licenses
- Management, operations, staff,
- Costs, economics
- Integration, cooperation with other resources, libraries, or services

An evaluation of a digital library, either in research or practice, could select from these elements, as to what to evaluate. In other words, an evaluation must specify clearly what elements are evaluated, with a full recognition of the emphasis. Every evaluation will leave something out. With the present state-of-the-knowledge, no evaluation can cover even the

majority of elements involved in a digital library, nor can it pretend to do so. Thus, there is no "evaluation of digital libraries." Possibly, there is only an evaluation of some of the elements in their construct.

Context for evaluation: At what level to evaluate?

Any evaluation is a tuplet between a selected element to be evaluated and a selected type of its performance. This leads to selection of a level of evaluation: What to concentrate on? Digital libraries, as other systems can be viewed and thus evaluated from a number of standpoints or levels. Each of these levels can be translated into a goal for evaluation.

A big dilemma and difficulty in evaluation is the selection of the level of objectives to address. Let us divide objectives, and thus evaluations, of a technical, computer-based system, such as a digital library into seven general classes or levels (of course, they are not mutually exclusive). The first three are more user-centered and the last three more system-centered with an interface in between. The performance questions for each level are indicated:

User-centered:

1 Social level: *How well does a digital library support the needs and demands, roles, and practices of a society or community?* This can be very hard to evaluate due to the diverse objectives of the society or community. Many complex variables are involved.
2 Institutional: *How well does a digital library support institutional or organizational mission and objectives? How well does it integrate with other institutional resources?* This is tied to institutional, organizational objectives. Also hard to evaluate for similar reasons.
3 Individual: *How well does a digital library (or given services) support information needs, tasks, activities of people as individual users*? It turns out that most evaluations tend to be on that level, probably because it is most direct and easiest to evaluate, though differences in perceptions can prove troublesome and it is not always easy to generalize to a larger population.

4 Interface: *How well does a given interface provide and support access, searching, navigation, browsing, and interaction with a digital library?* Questions can be asked in either the user or system direction or in both directions.

System-centered:

5 Engineering: *How well do hardware, networks, and related configuration perform?* These questions yield more replicable measures and are more easily generalizable than many user-centered approaches.
6 Processing: *How well do procedures, techniques, algorithms, operations … perform?* These are also very systematic though there may be variation due to differences in configuration, capacity, and other system variables.
7 Content: *How well is the collection or information resources, selected, represented, organized, structured … ?* Although this is also fairly systematic, the related question is how well for whom and for what purpose. Many systems are used in ways that their designers never intended.

Moreover, as mentioned, not only effectiveness but also efficiency or cost-effectiveness questions can be asked and contrasted at each level. Evaluation on one level rarely, if ever answer questions from another. For instance, evaluations of engineering or processing aspects of digital libraries say little about questions arising in evaluation of use. In real-life operations and applications of digital libraries at a number of levels are closely connected. In evaluations of digital libraries they are not. As yet, digital libraries are not evaluated on more than one level. This isolation of levels of evaluation could be considered a further and great challenge for all digital library evaluations.

Criteria for evaluation

Criteria for each level have to be determined. So far there is little agreement as to what should these criteria be. In the evaluations reviewed above a level was explicitly or implicitly chosen, and with it a set of criteria was used, as enumerated. Among the level chosen for evaluation most often was the individual level, as defined, and among the criteria the most prominent was usability.

Marchionini, Plaisant, and Komlodi (in press) at the outset of a chapter that among others addresses design and evaluation of digital libraries state:

> Digital libraries (DL) serve communities of people and are created and maintained by and for people. People and their information needs are central to all libraries, digital or otherwise. *All efforts to design, implement, and evaluate digital libraries must be rooted in the information needs, characteristics, and contexts of the people who will or may use those libraries*. (Emphasis in the original).

In this concept evaluation is squarely placed in the realm of user-centered levels, with an implicit if not explicit absence of system-centered levels. We disagree with the concept that evaluation must or should a priori be based on any one or a set of given levels, be they user- or system-centered. Evaluation can and should be performed at different levels, involving different objectives and related criteria. This issue has been visited and even vehemently argued a number of times in the debates about information retrieval (IR) design and evaluation. The conclusion about approaches to IR design and evaluation is valid for digital libraries as well:

> But, the issue is not whether we should have systems- OR human-centered approaches. The issue is even less of human- VERSUS systems-centered. *The issue is how to make human- AND systems-centered approaches work together.* (Emphasis in the original) (Saracevic, 1999).

For each of the levels criteria have to be developed and applied. For instance, there is nothing wrong in developing criteria for evaluation of the content level in relation to collection and asking questions such as: H*ow well does a given collection represent that which exist in a given domain or media? How timely it is? How well is it represented according to some standard?* The last question relates a digital library collection to some standards. These and like evaluative questions involve just that level and they are important for assessing a given collection by itself. Thus, not everything has to or should be centered in any one level or a given set of levels.

Adaptation

We enumerated a number of criteria that have been used in evaluation of digital libraries. Here we make suggestions about criteria that have been used in practice in related enterprises and that can be considered for adapting into criteria for digital library evaluation.

Libraries, information retrieval systems, and human-computer interfaces have been evaluated for a long time using numerous criteria. A good number of evaluation criteria for libraries were summarized by Lancaster (1993), for library and information services by Saracevic & Kantor (1997), for IR systems by Su (1992) and for interfaces by Shneiderman (1998). Buttenfield (1999) provides a framework for usability evaluation and criteria. From these and other sources we provide a short list of criteria that could be adapted for digital libraries:

Traditional library criteria :

- ← collection
  - purpose, scope, authority, coverage, currency, audience, cost, format, treatment, preservation ...
- ← information

accuracy, appropriateness, links, representation, uniqueness, comparability, presentation
…

- use
  accessibility, availability, searchability, usability …
- standards for a number of elements and process.

Traditional IR criteria

- relevance (leads to measures of precision and recall
- satisfaction, success …

Traditional human-computer interaction/interfaces criteria

- usability, functionality, effort …
- task appropriateness, failures

Conclusions

Digital libraries have exploded on the scene. Numerous research and practical efforts and large resources are expended on digital library research and practice. Evaluation is not, by and large, a part of these efforts. With few notable exceptions in either research or practice, digital library evaluation is not conspicuous. Despite these exceptions, digital library evaluation has yet to penetrate research, practice, or even debate. But it must be recognized that digital library evaluation is a complex and difficult undertaking. In this paper we enumerated the challenges facing digital library evaluation and suggested a conceptual a framework for evaluation, derived from the systems approach. A lot more has to be specified and agreed upon before digital library evaluation can be carried out in a consistent manner, a manner that would allow even for comparisons.

A significant point has been made in the opening statement on the web page of Digital Library Federation (1999):

> One of the great accomplishments of traditional libraries is that they are organized along similar lines. The individual who knows how to use one library in this country is likely to be able to use any other. Users have come to take this uniformity for granted in the print environment, but it is far from the norm in the digital environment.
> Digital resources now available through global networks are anything but organized. If digital collections created or stored at one library are to be available to others, there must be general agreement about the requirements for systems architecture, metadata, indexing, and retrieval. The development and adoption of common standards will require significant additional effort and exploration.

The evaluation of digital libraries should be also looking at and contributing to the gaining of uniformity for access and use across the landscape of digital libraries, which involves evaluation across a number of digital libraries not only single efforts. While it is way too early to set formal standards for digital libraries, and thereby freeze innovation, it is not too early to think about evaluation of factors and features contributing to uniformity, as an additional criteria.

Even if there is no visible movement in evaluation of digital library evaluation on a formal level, an informal evaluation of digital library efforts will proceed no matter what. Funders, users, the public, peers, technologists, experts, lay people, and everybody that gets in touch with results of digital library research or practice in any way do such evaluation. Such informal evaluations can be valid and reliable, but can also stray in significant ways and create erroneous perceptions and expectations of digital libraries. Thus, it is imperative that efforts in formal evaluation of digital libraries be enlarged and become an integral part of all research and practice, no matter what the challenge.

After all this is said of evaluation, a larger set of questions loom, questions to which we eluded in the Introduction:

At this early stage of digital library evolution isn't too early to concentrate on evaluation? Could early evaluation stifle innovation? Could it lead into different directions, such as concentrating on minutia of that which can be measured over the bigger picture? Could premature evaluation turn contraproductive?

If evaluation is taken rigidly, the answer to all of these questions is YES. But if taken in the spirit of evolution of digital libraries, then their evaluation should also be taken as an evolutionary enterprise. Evolution of evaluation should be treated as a necessary part of the larger evolution of digital libraries, and as that larger evolution it will have a part that ends in blind alleys and hopefully a much larger part that leads to successes. But, it is never to early to start thinking about it and to go on clarifying evaluation concepts and doing evaluation experiments. The paper has been written in that spirit.

References

Bishop, A. P. (1999). Document structure and digital libraries: How researchers mobilize information in journal articles. *Information Processing & Management, 35* (3) 255-279.

Bonn, M. S., Lougee, W. P., Mackie-Mason, J. K. &. Riveros, J. F. (1999). A Report on the PEAK Experiment. Context and Design. *D-Lib Magazine, 5* (6) URL: www.dlib.org.

Bourdieu, P. (1990). *The logic of practice.* Stanford, CA: Stanford University Press.

Borgman, C.L. (1999) What are digital libraries? Competing visions. *Information Processing & Management, 35* (3) 227-243.

Buttenfield, B (1999). Usability evaluation of digital libraries. *Science & Technology Libraries, 17* (3/4) 39-59.

Cornell University (1999). MESL technical report. URL: cidc.library.cornell.edu/gateway.htm

Digital Library Federation (1999) URL: www.clir.org/diglib/dlfhomepage.htm

Hill, L. L et al (2000) Alexandria Digital Library: User evaluation studies and system design *Journal of the American Society for Information Science, 51* (3) 246-259.

Lancaster, F.W. (1993). If you want to evaluate your library... Univ. of Illinois, Grad. School of Library and Information Science.

Lesk, M.E. (1997). *Practical digital libraries: Books, bytes, and bucks* San Francisco: Morgan Kaufman.

Licklider, J.C.R. (1965). *Libraries of the future*. Cambridge, MA: MIT Press.

Mackie-Mason, J.K., Riveros, J. F., Bonn, M. S., & Lougee, W.P. (1999). A Report on the PEAK Experiment. Usage and Economic Behavior. *D-Lib Magazine, 5* (7/8) URL: www.dlib.org.

Marchionini, G. & Crane, G. (1994). Evaluating hypermedia and learning: Methods and results from the Perseus Project. *ACM Transactions on Information Systems, 12* (1), 5-34.

Marchionini, G.; Plaisant, C.; & Komlodi, A. (in press) The people in digital libraries: Multifaceted approaches to assessing needs and impact. Chapter in Bishop, A. Buttenfield, B. & VanHouse, N. (Eds.) *Digital library use: Social practice in design and evaluation.* MIT Press. (http://ils.unc.edu/~march/revision.pdf)

Saracevic, T. & Kantor, P. (1997). Studying the value of library and information services. I. Establishing a theoretical framework. II. Methodology and Taxonomy. *Journal of the American Society for Information Science, 48* (6), 527-542, 543-563.

Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science, 50* (12), 1051-1063.

Schiff, L. R., Van House, N.A., & Butler, M.H. (1997). Understanding complex information environments: A social analysis of watershed planning. *Proceedings of the 2nd ACM International Conference on Digital Libraries,* 161-168.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human---computer interaction* (3nd ed.). Reading, MA: Addison-Wesley.

Su, L. T. Evaluation measures for interactive information retrieval. *Information Processing & Management, 28* (4), 503-516.