Digital Library Evaluation: Toward an Evolution of Concepts[1]

TEFKO SARACEVIC

ABSTRACT

WHILE THERE WERE MANY EFFORTS IN THE RESEARCH and practices of digital libraries, evaluation was not a conspicuous activity. It is well recognized that digital library evaluation is a complex and difficult undertaking. Challenges facing digital library evaluation are enumerated. A conceptual framework for evaluation is suggested. 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, included are suggestions for adaptation of criteria from related activities. The article 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 1980s, digital libraries (under various names) were barely a part of the landscape of librarianship, information science, or computer science. But just a decade later, by the end of the 1990s,

Tefko Saracevic, School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901-1071
LIBRARY TRENDS, Vol. 49, No. 3, Fall 2000, pp. 350-369

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 2000s.

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. The work of two communities, research and practice, are reviewed below. While they work and proceed mostly independently of each other, they can be considered as two ends of a spectrum, which as yet have not met in the middle. The research community, on one end of the spectrum, asks research questions directed toward future vision or visions of digital libraries, or rather of their various aspects and components, unrestricted by practice. On the other end of the spectrum, the practice community asks developmental and operational questions in real-life economic and institutional contexts, restrictions, and possibilities, concentrating on applications on the "market" end of the spectrum.

Large resources and efforts have been expended on digital library research and practice. There are many efforts, projects, and implementations, not only in the United States but in many other countries and on international levels as well. More are underway. Many exciting things are being done and explored. However, evaluation is more conspicuous by its absence (or just minimal presence) in the vast majority of published work on digital libraries, in either 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 than the rule. Why is that? Some speculations are:

- Perhaps it is too early in the evolution of digital libraries to attempt evaluation in any formal way. Evaluation at this stage of evolution may be premature and even dangerous because of possible stifling effects.
- At this stage, informal and anecdotal ways of evaluation suffice.
- Maybe evaluation is taken to be sufficient on a very basic technical level—the fact that something computes or that an electronic collection is searchable and accessible is sufficient as evaluation in itself.
- From a cynical perspective, we might suggest that the interest in evaluation is suppressed. Who would want to know about or demonstrate the actual performance?

- 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 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, I 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 The main purpose of this discussion is to address various conceptual and theoretical questions about the evaluation of digital libraries and to propose concepts and approaches believed to be appropriate toward their evaluation. The article is considered 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, the concentration here is, as mentioned, on the research and practice communities as being most closely evaluation bound. Each has a differing interpretation 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 evaluate digital libraries in the first place?

The distinction (and possible source of tension and lack of communication) between the two communities and approaches has been nicely illustrated by Rusbridge (1998) while contrasting two different approaches to digital libraries—i.e., the U.K. approach in the electronic libraries (eLib) program with the U.S. approach in the Digital Library Initiatives (DLI).

> The participants [at digital library conferences in the United States reflecting DLI] aimed (properly) to be innovative and free-thinking, leaving aside the constraints of existing practice The results are exciting and extraordinarily interesting, but it is very hard to determine how many of these ideas might be effectively deployed in real life situations It is notoriously difficult to transfer new technology from experiment to practice, but this is clearly harder the more distant the experimental context from real life
>
> By contrast, the eLib program characterised itself right from the start as "development" rather than research. . [The mission of Joint Information Systems Committee (JISC) funding the eLib projects] is *to stimulate and enable the cost effective exploitation of information systems and to provide a high quality national network infrastructure for the UK higher education and research communities*, in this context,

> JISC funds a number of development programs aimed at supporting universities by piloting the use of appropriate new technologies. Unlike the fundamental research characteristics of the NSF and similar agencies, JISC's projects are concentrated at the near-market practical application end of the spectrum. Both are needed. The eLib work is still research, despite a curious disdain for the word in some quarters.

Research Community

The research community, with most members having a background in computer science, concentrates on developmental research and experimentation in 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 United States, digital library research is guided, and even defined, through the projects supported by Digital Library Initiatives (DLI). The DLI 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. DLI-2 (1999-2003), funded by eight agencies, involves approximately sixty large and small projects. There are also large research digital library initiatives funded, among others, in the United Kingdom, Germany, Japan, Australia, New Zealand, and regionally by the European Union. This article concentrates on the efforts in the United States while recognizing the existence of many other efforts in many other countries and regions.

DLIs 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 DLI-2, 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 included 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 studied 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 ways to approach and implement 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 the University of California at Berkeley, as part of the evaluation, a series of interviews with intended users were conducted (Schiff, Van House, & Butler, 1997). They focused on situated actions, defined as "[action] performed by specific individuals in specific sociocultural 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. He 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 "mobilized the work . . . as they identify, retrieve, read and use material in articles of interest" (Bishop, 1999, pp 255-56). The criteria were *work* and *use of retrieved materials by users*. In another report by Bishop (1998), the criteria related to access were prominently investigated with results aimed at removing trivial and other barriers to access and use.

These three projects, and similar studies of user behavior related to digital libraries or to information in general, provide useful information, as pointed out by Bishop (1999), "[with] implications for user education and digital library system design" (p. 257). 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 not 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 resides in operational libraries, concentrates on building operational digital libraries, their mainte-

nance 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. Numerous types of collections and media are included and processed in many different ways Several are located in libraries, creating a hybrid library (combining a traditional and digital library), while 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* (http://www.dlib.org). The American Memory Project, pioneered by the Library of Congress, has provided a template for many other projects and is also among the earliest such projects to have paid attention to evaluation with criteria of use, usability, and a variety of technical aspects (http://memory loc gov/ammem/usereval.html and http://memory loc gov/ammem/ipirpt html).

Among the early and longest lasting evaluations of practical digital libraries is the evaluation of the Perseus Project, a corpus of multimedia materials and tools related to the ancient Greek world (http // www.perseus.tufts.edu). The mission of Perseus is to provide improved access to primary source materials related to the needs of students and faculty and to foster greater understanding of 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 identified: (1) learning, (2) teaching, (3) system (performance, interface, electronic publishing), and (4) 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 (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, Lougee, Mackie-Mason, & Riveros, 1999; Mackie-Mason, Riveros, Bonn, & Lougee, 1999) It is unique in that it 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 were offered to different groups), *pricing* (different models), and *revenues and costs* This project extended evaluation criteria and measures to economic factors 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—many questions have been asked of users, designers, developers, and operators to obtain evaluative impressions and assessments. Criteria in questions include functionality—i.e., 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, and skills. The evaluation was not formal, but it is interesting if for nothing else than 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—i.e., 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" (Jones, Gay, & Reiger, 1999). 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 later in this discussion.

Kilker and Gay (1998), in providing a framework for evaluation and applying that framework to a case study, expressed ideas that were similar to those of other studies at Cornell. The framework was the Social Construction of Technology (SCOT) theory, where the concentration is on examining varied conceptions held by "relevant social groups" involved in technology development and use. The approach is presented as an alternative to system and user-centered frameworks for study and evaluation. It recognizes that different audiences associated with a digital library (from designers to different groups of users) have different interpretations; they evaluate a digital library differently and use a different terminology. The criteria are: relevant social groups, interpretive flexibility (capabilities, responses), and mediation.

In a joint international undertaking, the National Science Foundation (primary sponsor of DLI research in the United States) and Joint Information Systems Committee (JISC) (primary sponsor of the eLib program in the United Kingdom) developed in 1999 a joint initiative or as they called it, "a hybrid process . . . in order to bring together the best elements of the styles of the two funding bodies." The idea is to fuse the two approaches, where the objective of JIST is "development of content or new technologies that would be widely applicable and not just of benefit to the participating institutions," while NSF's objective is "new research in the area of digital libraries, and the presence of new scientific ideas and methods." The efforts of funded projects are geared toward criteria showing "ability of the international partners to work together," and to combine research with practical development (Wiseman, Rusbridge, & Griffin, 1999). These criteria differ from others applied in either research or practice. However, from examination of abbreviated proposals, the funded projects under this international initiative have little in the way of evaluation built in.

There is still another practical concern that closely relates to evaluation. For over 100 years, ever since Melvil Dewey, library collections have been built and managed in relation to some established standards and policies. These provided criteria for traditional evaluation of collections. Not surprisingly, a number of libraries and library-related consortia that are in the process of developing or acquiring digital collections have also undertaken establishment of standards and policies for such collections. Okerson (1999) provides links to more than thirty library sites announcing their standards or policies for digital or electronic collections. In turn, these are incorporating, directly or indirectly, criteria for evaluation of digital collections and raising significant issues about the standards themselves and their use in evaluation. Most of the criteria incorporated are derived from traditional library collection criteria, as enumerated below—and they fit well. But, slowly, some additional criteria are emerging. Among them are strategic significance and availability of other distributed sources, such as found on the Internet or databases in the organization (e.g., on campus). Organizations not directly connected with libraries are also concerned with policies for digital collections and databases; good examples are the elaborate policies and criteria established by the Arts and Humanities Data Service (United Kingdom) (Beagrie & Greenstein, 1998). An important issue there is validation of sources included. It turns out that validation is a key problem in the use of Internet resources in general. The efforts of libraries to provide standards and criteria for their digital holdings that are then available (generally or to restricted audiences) over the Internet are establishing trust, validity, and authority for their own resources on the Internet, thus promoting access with user confidence, a highly important thing on the otherwise value-neutral Internet.

To summarize: this review includes representative efforts, primarily to illustrate criteria used. It does not claim to cover the entire subject But evaluation coverage generally is not large; few other evaluations were found in either research or practice This illustrates the point that there is a dearth of evaluation efforts in comparison to all efforts related to digital libraries.

Needed and Lacking for Digital Library Evaluation

The general questions in any and all evaluations are. Why evaluate? What to evaluate? How to evaluate? For whom to evaluate? There are many approaches to evaluation and to answering these questions. We must 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 in illuminating the social forces and effects. The economic approach is appropriate in accounting for economic factors, the political science approach for policy and political factors, and so on. Clearly, every approach has strengths and weaknesses, there is no one "best" approach It is naive to argue for a predominance of any given approach. The answer to the first question as to why to evaluate should serve as a base for selection of an appropriate approach or approaches.

However, here the concentration is on the systems approach only as the most widely practiced or suggested approach for evaluation of all kinds of information systems, including digital libraries, fully recognizing both its strengths and limitations. At the outset, the basic assumption of all systems approaches is that 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 the context of evaluation below

To establish a common vocabulary and concepts, a few standard definitions follow. A system can be considered as 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 components, interact to perform certain functions or processes to achieve given objectives. Furthermore, any system (digital libraries included) exists in an environment, or more precisely in a set of environments (which can also be thought of as systems, and some may think of this as contexts), and interacts with its environments. It is difficult, and even arbitrary, to set the boundaries of a system In the evaluation of digital libraries, as in the evaluation of any system or process, these difficult questions arise that clearly affect the results: Where does a digital library under evaluation begin to be evaluated? Where does it end? What are the boundaries? What to include? What to

exclude? On what environment or context to concentrate? This provides the questions for determining the construct of digital libraries, as discussed below.

In this context, evaluation means 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?
- efficiency At what cost (costs could be financial or involve time or effort)?
- a combination of these two (i.e., cost-effectiveness).

An evaluation has to specify which of these will be evaluated This discussion will primarily involve the evaluation of effectiveness with a realization that, during any evaluation of efficiency, 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 an evaluation, it has to meet certain requirements. It must involve selections and decisions related to the:

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 on criteria, measures, and methodologies for digital library evaluation, or 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 themselves, measurement, collection of statistics, or specification of metrics for digital libraries are not evaluation—they are quantitative or qualitative characterizations. 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 which includes all five requirements enumerated earlier.

A related view of evaluation is expressed by Marchionini, Plaisant, and 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 (p 2)

Construct: What Is a Digital Library?

What is there to evaluate? A simplistic answer is that whatever is called a "digital library" project is therefore considered a digital library, thus a construct candidate for evaluation. (This is derived from a certain philosophical stance whose metaphor is "Physics [or whatever field] is what a physicist does".) This is a pragmatic approach that has been at times applied to modeling a construct of a digital library; to some extent it even 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 a digital and distributed collection, the modeling and evaluation of digital libraries may, to some extent, parallel those related to physical libraries—at least initially. But (and this is a very important "but") digital libraries are also quite different and, in some functions, as for example in distribution and access, completely different from physical libraries. Thus, digital libraries also

require additional and new approaches to modeling of their constructs and thus to evaluation as well. Also, a digital library is much more than a collection of digitized texts and other objects. The challenge at the beginning of digital library evaluation is developing and applying these new modeling concepts to the specifics of what is meant by, and incorporated in, a "digital library."

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 into evaluation, answering the question at the start of this section. The question should be raised. Is this enough? I 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. (p. 230)

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; and
- community of users—their information needs and uses

In a newer text, Arms (2000) provides what he calls an "informal definition": "a digital library is a managed *collection of information*, with associated *services*, where the information is stored in *digital formats* and *accessible* over a *network*. The crucial part of this definition is that the *information is managed*" (p. 2, emphasis added). In this construct, the subjects for evaluation are italicized. The critical element added here is the aspect of management of the collection and information.

In the United States, the Digital Library 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" (DLF, September 17, 1999). The organization represents the practical community. After considerable work, DLF agreed on a "working definition of a digital library" representing a 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. (DLF, April 21, 1999)

This definition and conception is quite different from the one provided by Lesk (1997), Arms (2000), and even by Borgman (1999). Here the emphasis is on an 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 an integration. In a general way, the constructs or elements for evaluation of digital libraries are:

- digital collections, resources;
- selection, gathering, holdings, media;
- distribution, connections, links,
- organization, structure, storage;
- interpretation, representation, metadata;
- management;
- preservation, persistence;
- access;
- physical networks,
- 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, and
- integration, cooperation with other resources, libraries, or services.

An evaluation of a digital library, either in research or practice, could select what to evaluate from these elements. In other words, an evaluation must specify clearly what elements are evaluated with full recognition of the emphasis on what is included and what is excluded. Every evaluation leaves something out. With the present state of 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 element of its performance. This leads to selection of a level of evaluation: What to concentrate on? Digital libraries, like 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

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.

Institutional How well does a digital library support the 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.

Individual How well does a digital library (or given services) support information needs, tasks, activities of people as individual users or groups of

users with some strong commonalties? 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.

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

Engineering. How well do hardware, networks, and related configurations perform? These questions yield more replicable measures and are more easily generalizable than many user-centered approaches.

Processing. How well do procedures, techniques, algorithms, operations, and so on perform? These are also very systematic, though there may be variation due to differences in configuration, capacity, and other system variables.

Content How well is the collection or information resources selected, represented, organized, structured, and managed? Although this is also fairly systematic, the related questions are how well, for whom, and for what purpose?

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, answers questions from another. For instance, evaluations of engineering or processing aspects of digital libraries say little about questions arising in the evaluation of use. In real-life operations and applications of digital libraries, a number of levels are closely connected, but evaluations of digital libraries 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 greater challenge for all digital library evaluations. In addition, as a rule, many systems are used in ways that their designers never intended

Criteria for Evaluation

Criteria for each level have to be determined. So far there is little agreement as to what these criteria should be. In the evaluations reviewed above, a level was explicitly or implicitly chosen, and with it a set of criteria was used as enumerated 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 other things, 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* (p. 1, 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. I 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* (Saracevic, 1999, p. 1058, emphasis in the original)

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 the collection and asking questions such as: How well does a given collection represent that which exists in a given domain or medium? How timely is it? How well is it represented according to some standard? The last question relates a digital library collection to some standards. These and similar 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

A number of criteria used in the evaluation of digital libraries were enumerated. Next, suggestions are made 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 and 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, here is a short list of criteria that could, and even must, be adapted for digital libraries.

Traditional library criteria

- collection: purpose, subject, scope, authority, coverage, currency, audience, cost, format, treatment, preservation, persistence;
- information: accuracy, appropriateness, links, representation, uniqueness, comparability, presentation, timeliness, ownership;
- use: accessibility, availability, searchability, usability; and
- standards for a number of elements and processes

Traditional IR criteria

- relevance (leads to measures of precision and recall),
- satisfaction, success; and
- index, search, output features.

Traditional human-computer interaction/interfaces criteria

- usability, functionality, effort,
- task appropriateness, failures,
- connectivity, reliability;
- design features;
- navigation, browsing; and
- services, help.

Conclusion

Digital libraries have exploded onto 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. This article discusses the challenges facing digital library evaluation and suggests a conceptual framework for evaluation derived from the systems approach. Much 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 the 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 also be 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 and 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 criterion. A further and critical issue for evaluation is persistence. An important feature of many traditional libraries is that their collections are preserved over time—they persist. An important feature of digital collections is a potential lack of persistence. Libraries have no control whatsoever over persistence of digital journals, indexes, and the like for which they have licensed access for a time. Publishers may go out of business, as many do, or they may change direction into some other line, as also many do, and thus the sources under their control will vanish. Digital journals, such as *D-Lib Magazine* copiously cited here, may vanish after their funding runs out. Persistence may become one of the most important criteria for digital libraries.

Even if there is no visible movement in the evolution of digital library evaluation on a formal level, an informal evaluation of digital library efforts will proceed by funders, users, the public, peers, technologists, experts, lay people, and anybody that is involved with the results of digital library research or practice in any way. 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 I alluded to in the introductory comments. "At this early stage of digital library evolution is it 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 it is hoped a much larger part that leads to successes. But, it is never too early to start thinking about it and to go on clarifying

evaluation concepts and doing evaluation experiments. This article has been written in that spirit.

The ultimate evaluation of digital libraries will be in relation to the transformation of their context, the same as of evaluation of libraries throughout history. Digital libraries provide for an interaction among people, human knowledge, organizations, and technology. The ultimate question for evaluation is How are digital libraries transforming research, education, learning, and living? At this stage, we don't have the answers, but we have indications that significant transformations are indeed taking place.

NOTE

[1] This paper is substantially based on Saracevic, T., & Covi, L. (2000) Challenges to digital library evaluation *Proceedings of the American Society for Information Science, 37*

REFERENCES

Arms, W. Y. (2000) *Digital libraries* Cambridge, MA: MIT Press

Beagrie, N., & Greenstein, D. (1998) *Managing digital collections: Arts and Humanities Data Service (AHDS) policies, standards and practices: Evaluation criteria for evaluation* Retrieved September 29, 2000 from the World Wide Web: http://ahds.ac.uk/public/srg.html#criteria

Bishop, A. P. (1998) Measuring access, use, and success in digital libraries *Journal of Electronic Publishing, 4*(2) Retrieved September 29, 2000 from the World Wide Web: http://www.press.umich.edu/jep/04-02/bishop.html

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) Retrieved September 29, 2000 from the World Wide Web: http://www.dlib.org/dlib/june99/06bonn.html

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* Retrieved September 29, 2000 from the World Wide Web: http://cidc.library.cornell.edu/gateway.htm

Digital Library Federation (1999, April 21) *A working definition of digital library* Retrieved September 29, 2000 from the World Wide Web: http://www.clir.org/diglib/dldefinition.htm

Digital Library Federation (1999, September 17) *DLF Home* Retrieved September 29, 2000 from the World Wide Web: http://www.clir.org/diglib/oldhome.htm

Hill, L. L., Carver, L., Larsgaard, M., Dolin, R., Smith, T. R., Frew, J., & Rae, M. A. (2000) Alexandria Digital Library: User evaluation studies and system design *Journal of the American Society for Information Science, 51*(3), 246-259

Jones, M. L. W., Gay, G. K., & Rieger, R. H. (1999) Project Soup: Comparing evaluations of digital collection efforts *D-Lib Magazine, 5*(11) Retrieved September 29, 2000 from the World Wide Web: http://www.dlib.org/dlib/november99/11jones.html

Kilker, J., & Gay, G. (1998) The social construction of a digital library: A case study examining implications for evaluation *Information Technology and Libraries, 17*(2), 60-70

Lancaster, F. W. (1993) *If you want to evaluate your library* (2d ed.). Urbana-Champaign: University of Illinois, Graduate 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). Retrieved September 29, 2000 from the World Wide Web: http://www.dlib.org/dlib/july99/mackie-mason/07mackie-mason.html

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. In A. Bishop, B. Buttenfield, & N. VanHouse (Eds.), *Digital library use: Social practice in design and evaluation*. Cambridge, MA: MIT Press. Retrieved September 29, 2000 from the World Wide Web: http://ils.unc.edu/~march/revision.pdf

Okerson, A. (1999). *Electronic collections development*. Retrieved September 29, 2000 from the World Wide Web: http://www.library.yale.edu/~okerson/ecd.html

Rusbridge, C. (1998). Towards the hybrid library. *D-Lib Magazine, 6*(7/8). Retrieved September 29, 2000 from the World Wide Web: http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html

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.

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

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management, 28*(4), 503-516.

Wiseman, N., Rusbridge, C., & Griffin, S. M. (1999). The joint NSF/JISC International Digital Libraries Initiative. *D-Lib Magazine, 5*(6). Retrieved September 29, 2000 from the World Wide Web: http://www.dlib.org/dlib/june99/06wiseman.html