Fourth DELOS Workshop. Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics

Hungarian Academy of Sciences
Computer and Automation Research Institute (MTA SZTAKI)
Budapest, Hungary
6-7 June 2002

FINAL REPORT TO
NATIONAL SCIENCE FOUNDATION
COMPUTER AND INFORMATION SCIENCE DIRECTORATE
INFORMATION AND INTELLIGENT SYSTEMS DIVISION
Digital Libraries Program
Stephen Griffin, Program Officer

Grant IIS-0225626

Christine L. Borgman, Principal Investigator
Professor & Presidential Chair in Information Studies
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1520
cborgman@ucla.edu

Acknowledgements

The draft of this report was circulated for comment to all participants in the workshop. This final report incorporates substantial contributions and editorial comments from Ingeborg T. Sølvberg, European Program Co-chair of the Workshop and U.S. workshop participants Nicholas Belkin, Ann Bishop, Ronald Larsen, Javed Mostafa, and Robert Sandusky.

The success of this workshop was due to the contributions of many people, especially my program co-chair, Ingeborg T. Sølvberg, Laszlo Kovacs, General Chair, and András Micsik and Zoltan Tóth, who handled the website and the local arrangements.

This report and all other documents from the workshop, including agenda, papers, breakout group reports, and list of participants, are available on the workshop website:

http://www.sztaki.hu/conferences/deval/presentations.html

Table of Contents

I. WORKSHOP GOALS, OUTCOMES, AND RECOMMENDATIONS

1. Goals of the Workshop

This grant funded U.S. participation in the joint EU-US *Workshop on Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics.* The workshop was initiated by EU participants via the DELOS Network of Excellence for Digital Libraries, which is a framework for international cooperation on research activities and research agendas in the digital library domain. One of DELOS' most important activities is to sponsor workshops that bring together participants from multiple countries that are working in this inherently international and interdisciplinary research area. DELOS and the U.S. National Science Foundation have a long record of cooperation for joint efforts in the digital library arena.

Digital libraries can be viewed from a number of perspectives. They can be new forms of information institutions, multimedia information retrieval systems, or information systems that support the creation, use, and searching of digital content. Digital libraries are not ends in themselves; rather, they are enabling technologies for digital asset management, electronic commerce, electronic publishing, teaching and learning, and other activities. Accordingly, digital libraries need to be evaluated in the context of specific applications. The methods and metrics for evaluating digital libraries will vary by whether they are viewed as institutions, as information systems, as new technologies, or as new services.

The DL research communities need large test beds (collections and testing mechanisms) as a means to evaluate new concepts. Research results are most valuable when they are compared with other approaches and validated against other sets of data. Evaluations may involve users, collections, or systems.

László Kovács of the Hungarian Academy of Sciences was General Chair and host for the workshop. The workshop program was co-chaired by Christine L. Borgman of UCLA and Ingeborg T. Sølvberg of NTNU, Norway. This workshop brought together researchers and practitioners whose work includes evaluation of digital libraries in a variety of environments, using a variety of methods. Papers were invited that focus on generalizable metrics or on methods and measures specific to individual digital library contexts. These included, but were not limited to, education, publishing, cultural heritage, science and technology, medicine, sound, and images. Papers on context-specific evaluation methods provided background on the application, explanations of how and why evaluation is tailored, and the expected use of results (e.g., to improve learning, improve retrieval, improve navigation facilities). Some papers indicated how their approaches might be adapted to other contexts.

We especially invited DL evaluation papers that address organizational contexts, creation and use of content, and information retrieval. Thus, this workshop brought together researchers from different fields, such as library and information science, publishing, computer science, and content provision to exchange their ideas about DL evaluation.

DELOS funded some of the participation for European DELOS members who are actively involved in the workshop (accepted papers, speaking on a panel, serving on the program committee).

The funding from NSF provided travel support for the six participants from the U.S.
U.S. Members of the Program Committee:

Nicholas Belkin, Rutgers University
Ann Bishop, University of Illinois
Christine Borgman, UCLA
Ronald Larsen, University of Maryland (now at University of Pittsburgh)
Clifford Lynch, Coalition for Networked Information

Two invited speakers were chosen by the joint European-U.S. program committee. The six NSF funded participants were four members of the program committee, all of whom presented papers (Belkin, Borgman, Bishop, Larsen), a PhD student whose submitted paper was accepted for presentation (Robert Sandusky, University of Illinois), and one invited scholar working in an important area of DL evaluation not otherwise represented at the workshop (Javed Mostafa, Indiana University). Dr. Mostafa is conducting DL research on information filtering and bioinformatics with funding from NSF and on multimedia with funded by the U.S. Institute for Museum and Library Services. He chaired a session at the workshop and contributed actively to the discussions.

Larsen (this volume) succinctly summarizes the issues that led to this workshop:

> "A major challenge for digital library evaluators is to find relatively non-intrusive, low cost means of capturing appropriate data to expose and explore the dynamics underlying the use of digital libraries. This is the challenge for the DELOS Workshop on Evaluation of Digital Libraries."

2. Why Is Digital Library Evaluation Important?

Digital libraries have become an essential foundation for areas as diverse as electronic publishing and strategic defense, and serve as a primary means to deliver content for scholarship, commerce, cultural heritage, and education (including the National Science Foundation's NSDL program). Networked information systems are now an ubiquitous component of business, commerce, community, and education. Despite these advances, we have little understanding of the effectiveness of digital library systems and services in supporting these essential aspects of daily life in the 21st century.

Digital libraries support specific activities in specific contexts – classroom instruction, distance learning, digital asset management, scholarship, virtual museums, and so on. Digital libraries need to be evaluated as systems and as services to determine how useful, usable, and economical they are and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible. Consistent evaluation methods also will enable comparison between systems and services.

Evaluation research can be a highly applied form of investigation, or it can test theory. Evaluation research is particularly useful for studying aspects of communication technologies such as interactivity, adoption, use, implementation, and social impacts (Rogers, 1986). Evaluation itself can be cost effective, particularly in areas of usability. Even a small amount of usability evaluation in the development of information systems can pay for itself several times over in cost savings from lost productivity (Computer Science and Telecommunications Board 1997; Landauer 1995; Nielsen 1993; Sawyer, Flanders, and Wixon 1996).

2. What is Evaluation?

Evaluation is a general term that includes various aspects of performance measurement and assessment. Activities include laboratory experiments, regional, national, and international surveys or quasi-experiments, time-series analyses, online monitoring of user-system interactions, observation of use, and other forms of data collection. Evaluation has a long history in fields such as computer science, education, communication, health, and criminal justice. The effectiveness of interventions such as new teaching methods, management practices, and policy can be assessed (Burstein & Freeman, 1985; Rogers, 1986; Williams, Rice, & Rogers, 1988). Digital libraries can be viewed and evaluated as interventions in these fields, drawing upon methods typically used to assess the outcomes of programs and services.

In computer science, evaluation should be a continuous process throughout the life cycle of a system. The quality assessment process should distinguish between goals and means, as in the framework proposed by Lindland, Sindre, and Sølvberg (1994). In human-computer interaction, measures include time to learn, error rates, efficiency, memorability, and satisfaction (Nielsen, 1993; Shneiderman, 1998). Systems can be benchmarked for aspects of performance, using quantitative measures specific to applications, such as recall and precision measures of information retrieval. Aspects such as verification, validation, and quality assurance are based upon systems, technical and user requirements. Some useful definitions are these:

> *Quality assurance*: a planned and systematic pattern of all actions necessary to provide adequate confidence that a product (here software) conforms to established technical requirements (IEEE Software Engineering Standards Collection, 1991).
>
> *Verification:* the process of determining whether the products of a given development phase satisfy the requirements established during the previous phase (Thayer & Dorfmann, 1990).
>
> *Validation:* determining the correctness of the final program or software produced from a development project with respect to user's needs and requirements (Thayer & Dorfmann, 1990).

Evaluation methods should meet accepted norms for scientific rigor in the domain of study. In the social sciences, methods should be valid (be a "true" measurement of the quality or concept under study) and reliable (the same measure should achieve the same result at multiple times). Kirk and Miller (1986, page 80) offer succinct definitions of these concepts:

> *Reliability:* the extent to which the same observational procedure in the same context yields the same information.
> *Validity:* The quality of fit between an observation and the basis on which it is made.

At least four types of evaluation are relevant to digital libraries:

Formative evaluation begins at the initial stages of a development project to establish baselines on current operations, set goals, and determine desired outcomes. Such evaluation is usually driven by context and project-specific goals.

Summative evaluation takes place at the end of a project to determine if the intended goals were met. Goals and outcomes must be compared to initial states, so formative evaluation generally precedes summative evaluation.

Iterative evaluation takes place throughout a project, beginning in the earliest design and development stages. Interim stages of design are assessed in comparison to design goals and desired outcomes, and the results inform the next stages of design. Iterative approaches encourage designers to set measurable goals at the beginning of a project and provide opportunities to re-assess goals throughout the development process.

Comparative evaluation requires standardized measures that can be compared across systems. Communities can identify and validate measures. If such measures are implemented in a consistent manner, they enable comparisons between systems. Test beds are another way to compare measures and to compare performance of different functions and algorithms.

3. Prior U.S. and E.U. Research Activities on DL Evaluation

The *Workshop on Evaluation of Digital Libraries*, jointly funded by the European Union (via the DELOS Network of Excellence) and by the National Science Foundation, was preceded by many related activities in the United States, Europe, and Asia. We briefly summarize the prior U.S. activities and the prior European activities on evaluation of digital libraries. Some of these were joint U.S. – European efforts.

United States Activities on Evaluation of Digital Libraries

As part of the Digital Library Initiative, DARPA and NSF funded the Dlib Test Suite and Metrics Working Group. Ronald Larsen reported on the results of those efforts at the workshop (Larsen, this volume). The test suite provided DL researchers with access to large, standardized sets of data for quantitative and qualitative research in a distributed environment. The metrics working group considered evaluation issues in the system, user, and content domains. Their objective was to establish a rigorous set of metrics for comparative evaluation. They also identified a set of scenario-based challenge problems.

Digital libraries are difficult to evaluate due to their richness, complexity, and variety of uses and users. During the first Digital Library Initiative, NSF funded a workshop on *Social Aspects of Digital Libraries* (Borgman et al, 1996). The need for evaluation methods and metrics was among the key findings of that workshop. Saracevic (2000) also speculated that evaluation methods and models are insufficiently developed to address the complexity of digital library services. Some progress is being made, as evidenced by a special issue of *Library Trends* on "Assessing Digital Library Services" (Peters, 2000) and by a forthcoming book on the evaluation of digital libraries (Bishop, Van House, and Buttenfield, in press).

Projects funded under the first Digital Library Initiative included some evaluation components, notably the Alexandria Digital Library Project at the University of California, Santa Barbara (Buttenfield, 1999; Hill et al, 2000) and DeLiver at the University of Illinois (Bishop, 1998, 1999; Bishop, Neumann, Star, Merkel, Ignacio, & Sandusky, 2000). Phase 2 of the DLI included yet more evaluation components, such as the Alexandria Digital Earth Prototype (Borgman, et al., 2000; Gilliland-Swetland & Leazer, 2001; Leazer, Gilliland-Swetland, & Borgman, 2000 Leazer, Gilliland-Swetland, Borgman, & Mayer, 2000), and research with children at Maryland (Druin, et al., 2001). A recent study funded by DARPA found that developers of information systems could implement evaluation efforts successfully by sharing expertise among projects (Morse, 2002).

Other U.S. entities are beginning to fund DL evaluation efforts, such as the Institute for Museum and Library Services (e.g., Bishop, Mehra, Bazzell, & Smith, 2000, 2001). The Andrew W. Mellon Foundation provided funding for a workshop on evaluation frameworks for DLs of music (Downie, 2002). In the academic library community, efforts are underway to establish metrics for networked information services (Shim, 2001). Other private foundations are beginning to fund assessments of digital information resources that people use in everyday life (e.g., Berland et al, 2001). The proceedings of the first two *Joint Conferences on Digital Libraries* include a number of papers on digital library evaluation (Fox & Borgman, 2001; Marchionini & Hersh, 2002).

The efforts to date have been effective in establishing the need for evaluation of DLs, in identifying some of the areas most likely to be productive, and in demonstrating the effectiveness of small-scale evaluation efforts. However, they also showed the limitations of current evaluation efforts. The test suite was not as effective in engaging researchers in evaluation efforts as hoped. The metrics working group examined metrics and suggested scenarios, but did not validate them nor did they address research methods. (Larsen, this volume). Context-dependent evaluation efforts, while effective, remain hand-crafted and expensive (Belkin, Borgman, Bishop, this volume). One of the major problems in accomplishing evaluation is the lack of expertise and resources.

European Activities on Evaluation of Digital Libraries

Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics was the fourth in the DELOS Workshops series. The workshop was initiated and organized by the DELOS Working Group 2.1, which is responsible for providing a Digital Library Evaluation Forum and a Digital Library Test Suite.

The three previous DELOS workshops are:

- "Information Seeking, Searching and Querying in Digital Libraries", 11-12 December 2000, Zurich, Switzerland.

- "Personalisation and Recommender Systems in Digital Libraries", 18-20 June 2001, Dublin, Ireland.

- "Interoperability and Mediation in Heterogeneous Digital Libraries", 8-9 September 2001, Darmstadt, Germany.

On-line copies of the Proceedings of the DELOS Workshops are available on the ERCIM web-server: http://www.ercim.org/publication/workshop_reports.html Printed copies can be ordered from the same site.

DELOS Working Group 2: Evaluation

DELOS Network of Excellence (DELOS NoE) (www.delos-noe.org) aims at providing a *Digital Library Evaluation Forum* and *Digital Library Test Suites*. Three activities have been conducted during the years 1999-2002; the Cross-Language Evaluation Forum, the Metalibrary and DL Schema, and INEX: Testbed for XML retrieval. DELOS is an activity within the European Research Consortium for Informatics and Mathematics (ERCIM) (http://www.ercim.org)

Cross-Language Evaluation Forum (CLEF)

The volumes of information available over the global networks in languages other than English are increasing much faster than is the corpus of English language content. The user community for non-English language sources is creating enormous pressure for the development of systems that provide access to information without language or cultural barriers. For these reasons, Cross-Language Information Retrieval (CLIR) is a key topic for the Digital Library domain. However, the development of CLIR systems implies the need for suitable methodologies and tools to evaluate system performance.

The Cross-Language Evaluation Forum (CLEF) supports global digital library applications by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

The primary goal of CLEF to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness on the global marketplace. CLEF has arranged three Workshops, all organized in conjunction with the ECDL conferences. Proceedings are, or will be, published by Springer, in the series Lecture Notes in Computer Science (Peters, Braschler, Gonzalo, & Kluck, 2002; Peters, 2000). CLEF has now obtained independent funding from the EU Commission's IST programme. However, close contacts with DELOS will be maintained. Workshops will continue to be organized in conjunction with the ECDL conference series.

Further information on CLEF is available at http://clef.iei.pi.cnr.it/.

Metalibrary and Digital Library Schema

Several excellent collections have been or will be created with EU funding. To maximize the benefit of this work and in collaboration with their owners, the DELOS Network of Excellence' goal is to undertake the task of promoting the creation, maintenance and operation of a Test Suite, which will make test beds available to other researchers. This action will improve the efficiency of research since the Test Suite will provide all researchers with readily available resources for testing purposes.

A generic classification and evaluation scheme for digital libraries was developed (Fuhr et.al. 2001). The scheme is based upon the belief that evaluation of DLs should include a broad view of subject areas. The description scheme has four major dimensions: data/collection, system/technology, users, and usage. For each of these dimensions the major attributes are described. Overall, the original scheme has proven to be very useful; only a few suggestions for modifications of the scheme have been received.

A questionnaire about digital library test collections was created to gather input for the design of the European Digital Library Test Suite, and is being continued by setting up a DL Metalibrary. The DL Metalibrary has now 35 entries describing possible test beds worldwide.

The planned Test Suite will provide collections for comparative and quantitative experiments. This issue was the focus of a workshop session on "Evaluation of Digital Libraries" held on 8 February 2001 at the First EU-DL All Projects Concertation meeting in Luxembourg. (http://www.iei.pi.cnr.it/DELOS/delos2/International/sessionB)

Unfortunately, the effort to create, maintain and operate a DL Test Suite is currently postponed due to lack of funding.

Further information and continuing reports on the DL schema and metalibrary can be found at: http://www.sztaki.hu/delos_wg21/.

INEX: Testbed for XML retrieval

Many digital library documents are in XML format. DELOS supported the idea of creating a test bed for XML retrieval. INEX (Initiative for the Evaluation of XML Retrieval) is a coordinated effort of the University of Dortmund and the Queen Mary University of London, which received additional funding from DELOS for performing the infrastructure work of the evaluation process (Fuhr, Gövert, Kazai, Lalmas, 2002). As a collection, INEX uses about 12,000 journal articles from the field of computer science published by IEEE-CS during the years 1995-2001. After the call for participation in March 2002, 49 groups signed up. Finally, 25 groups were participating actively (several groups had to give up due to the complexity of the task), performing retrieval for 60 topics. Unlike TREC and CLEF, the relevance judgements in INEX are performed by the participating groups (each group has to judge document relevance for about 2 topics). The results of the initiative will be presented at a final workshop at Schloss Dagstuhl (Germany) in December 2002.

Further information on INEX is available at http://qmir.dcs.qmw.ac.uk/INEX/.

4. Workshop Themes

The workshop papers addressed a wide range of topics in the evaluation of digital libraries. For practical purposes, we organized the workshop sessions into four topical areas (users and user interfaces, evaluation in context, metrics and test beds, and evaluation of DL services and scalability), prefaced by a session of reports from prior working groups on DL evaluation. The four topic areas also provided a starting point for organizing the breakout groups.

Two broad themes emerged from the papers, breakout groups, and the rich plenary discussions that took place at the workshop: the complementary needs for (1) metrics and test beds and for (2) evaluation in the context of specific digital library applications.

Metrics and Test beds

The digital library community needs benchmarks for comparison between systems and services. Standards are required for DL architecture and operations if we are to achieve interoperability between systems and services. Similarly, the ability to scale DLs to full operational status, with ever-larger collections, will depend upon workable standards and interoperability.

Constructing test beds is beyond the capability of individual investigators or research teams. Test beds could be built specifically for comparing DL functions and services, as in the TREC experiments and a similar initiative in Japan for comparing cross-language information retrieval (Kando, this volume). They can also be organized as a collaboration among research groups, such as the Dlib test suite project (Larsen, this volume) and the DELOS Network of Excellence Working Group on Evaluation. The workshop breakout group on Metrics and Test beds (Sølvberg, Chair, this volume) sketched a model for test bed requirements.

We also need a set of metrics for comparing digital libraries. While the implementation of metrics may vary considerably by context, as discussed below, establishing a common set of metrics is essential for the reliability of DL evaluation. The Dlib metrics working group earlier identified 7 dimensions for DL metrics (Larsen, this volume). The DELOS Evaluation Forum defined a generic classification and evaluation scheme consisting of four major dimensions each with major attributes and metrics (Fuhr et al, 2001; Mabe, this volume). The workshop breakout group on Metrics and Test beds (Sølvberg, Chair, this volume) identified other metrics and some criteria for establishing metrics. The breakout group defined a Test bed as *a digital library and an evaluation goal.*

More detail on metrics and test beds can be found in the papers in the sessions on Background (Larsen, Mabe, Kando), Metrics and Test beds (Peters, Fuhr, Sandusky, Monch), and Services and Scalability (Abbattista et al, Griffiths & Fisher, Banwell), and in the breakout group report on Metrics and Test beds (Sølvberg, Chair).

Context and Applications

Test beds and metrics are most effective when problems are well understood. However, digital libraries are a new technology that is just beginning to move from research to practice and from prototypes to operational systems and services. As DLs are implemented, people gradually adopt and adapt them as part of their information practices. These behaviors are evolving rapidly, along with the implementation of systems. Thus, now is an excellent time to be studying uses, users, and usability of digital libraries and other aspects of DL context.

Context has a variety of aspects, including goals and tasks, socio-cultural milieu, and environment (breakout group on Evaluation in Context, Belkin, Chair), and these aspects must be considered with respect to research questions and methods. That breakout group identified 5 classes of research questions associated with context and evaluation of DLs, and suggested appropriate methods to address those questions. Evaluation of users and interfaces also must take place in a context, so that aspects such as domain, language, culture, format (text, audio, visual, etc.) can be assessed. The breakout group on Users and Interfaces (Borgman, Chair) also identified research questions and methods for studying DLs in context. The latter group proposed some criteria for determining the "best" research questions and methods, such as the cost of evaluation, cost-benefit of evaluation, adaptability of methods, sharability of methods, instruments, and test beds, and validity and reliability. Both groups concluded that evaluation can serve many different goals, and that the effectiveness of evaluation metrics and methods must be goal-specific. Methods and metrics to evaluate usability are unlikely to yield cost-benefit data and vice versa, for example.

Because digital libraries serve such a rich variety of content to a such a vast array of user populations, most DL evaluation to date has been specific to a context. Methods are often handcrafted and are time consuming to develop and deploy. We need more experience with context-specific evaluation methods to produce methods that can be applied more easily in new contexts. For example, methods used in the context of developing the capabilities and improving the life conditions of marginalized groups, such as participatory action research, can be applied to the evaluation of digital libraries (Bishop, et al., 2001; Freire, 2002; Harris & Weiner, 1996; Whitmore, 1998; Whyte, 1991). We also need to conduct evaluation in a wide variety of contexts to determine the commonalities and differences among digital libraries along various dimensions. Thus, research on digital libraries in specific contexts will lead to better metrics and methods that can be applied across digital library systems and services.

Further discussion of digital library evaluation in context is presented in the reports of the breakout groups on Evaluation in Context (Belkin, Chair) and on Evaluating Digital Library Users and Interfaces (Borgman, Chair) and in the papers in sessions on Users and User Interfaces (Ford et al, Sfakakis & Kapidakis), Evaluation in Context (Belkin, Bishop & Bruce, Borgman, Evans et al), and Evaluation of Services and Scalability (Abbattista et al, Friffiths & Fisher, Banwell). The report on the Dlib Metrics and Test bed efforts (Larsen, this volume) also addressed metrics that could be applied across contexts.

5. Workshop Recommendations

We allowed a substantial amount of time for discussion in the plenary sessions of the workshop, the breakout groups on each of the two days, and over meals. After the end of the workshop, the U.S. participants met with the DELOS Working Group members to discuss the outcomes and recommendations. The recommendations here are compiled from reports of the four breakout groups, from notes taken by U.S. participants in the plenary sessions and post-workshop discussion, and from subsequent commentary on the draft report.

Breakout Group Recommendations

Each of the theme breakout groups (Test Beds and Metrics, Evaluation in Context, Users and User Interfaces) identified research agendas for their areas, and we devoted an additional breakout group on the second day of the workshop to Next Generation Initiatives (Larsen, Chair). The latter group considered European Union efforts such as the 5th and 6th frameworks, U.S. efforts such as the NSF-led, multi-agency Digital Libraries Initiatives, the National Science Digital Library, and TREC workshops, and Asian efforts in digital libraries and in cross-language information retrieval. Digital libraries is a very successful arena for international cooperation, with many joint efforts among European, U.S., and Asian researchers. All three communities were represented at the workshop.

Evaluation of digital libraries also will require substantial international cooperation due to the distributed nature of digital libraries, the diversity of content and services, the need for multi-lingual content and user interfaces, and the variety of contexts. Also noted was the need to conduct research not only in academic environments but also in business, community, and social settings. Digital libraries constructed by community organizations such as public libraries, community networks, and hospitals are examples of important but under-studied environments.

The breakout group on Evaluation in Context proposed that the research agenda for evaluation of digital libraries be generalized to consider DLs as a class of "Complex Networked Information Systems" (CNIS). In this respect, they proposed four significant research areas: toolkits for CNIS evaluation, test bed of user interactions with CNIS, comparison of multiple aspects of CNIS, and means to incorporate users into the evaluation cycle.

The breakout group on Next Generation Initiatives set DL evaluation in a yet larger context, noting the relationships between digital libraries, grid computing, semantic web, and agent-based computing. These communities each need useful metrics and test beds and have similar challenges of critical mass and cooperation in developing them. A wide array of studies is required to understand how systems and users perform in different contexts. Many research challenges cross these four areas, including scaling, interoperability, usability, and services. The group concluded that in an era of global information systems and services, international collaboration is a technological necessity.

General Recommendations

In a 3-hour session following the workshop, U.S. participants and members of the DELOS working group on evaluation of digital libraries outlined the dimensions of DL evaluation. We summarize the dimensions as follows (based on notes by U.S. participants; this is not an official report of the joint group):

Requirements for the Evaluation of Digital Libraries:

As a community of research and practice, we lack:

–common evaluation resources (e.g. test beds, toolkits)

–metrics that are applicable to the DL situation as a whole, and across the different contexts in which DLs occur.

–methods for establishing relationships between users and uses of DLs.

As a community of research and practice, we need to develop

–means to take context into account in the design and evaluation of DLs.

–task- and situation-based measures and methods for evaluation of DLs.

–techniques for measuring impact of DLs.

–means to determine relationships among current evaluation measures, as a basis to develop more general measures and methods.

–relationships between the results of studies of user behavior and needs, and DL system design.

Identifying the above dimensions led to a preliminary research agenda for evaluation of digital libraries. Many, if not most, research projects resulting from this agenda would benefit by joint investigation between US and EU collaborators. Some potential avenues of research that should be explored with the goal of developing specific research project proposals are these:

1. Development of a DL evaluation infrastructure. This could include joint development of common standards for collection of records of interaction in DLs, and subsequent establishment of an institution which would collect such records from different groups, and put them together into a general resource for use as a test bed by others. It seems likely that such a project/program would be best implemented with international support.

2. The development of DL-specific evaluation metrics. Most metrics used to evaluate DLs to date have been derived from other contexts, for instance, information retrieval, or databases, or human-computer interaction. Very few of these even begin to reflect the totality of the DL situation. An international working group to develop and test new metrics, and to establish relationships amongst the different metrics, would lead to an immense improvement in our ability to evaluate DLs. This could be accomplished through funding relevant projects in different contexts/countries, and funding collaboration amongst them.

3. As an aspect of point 2, above, techniques for contextual evaluation should be investigated. Evaluation metrics and methods for DLs cannot be only general; they must also be sensitive to the context in which they are applied, and must take account of the context in order to come to valid conclusions. Projects that explicitly aim at developing context-dependent and context-sensitive evaluation techniques should be strongly encouraged.

Subsequent online discussion among the U.S. participants yielded a broad recommendation that incorporates many of the issues raised in this report. One of the inherent difficulties is accomplishing DL evaluation is the formation of a DL evaluation community, per se. The workshop itself quickly became a community of DL evaluation researchers and practitioners. Together, we drew upon our vast and diverse expertise to produce the analysis and recommendations in this report, in just two short days. Thus, the workshop is evidence that the basis for such a community exists.

Much of what we need is a set of resources and mechanisms to nurture and support such a community. TREC (the Text Retrieval Evaluation Conference) is a positive model of what can be accomplished by a dedicated community. The TREC model is valuable for the test corpora and metrics, but the real strength behind its success is the formation of a community with effective discussion lists, a well-maintained web site to distribute information, an annual gathering, and paid "community support staff" who spend a great deal of time conducting TREC-related activities that sustain the community.

Given the importance of DL evaluation, it has the potential to draw the attention of diverse communities engaged in DL creation - humanists, artists, social scientists, astronomers, bioinformaticians, geographers, etc. DL evaluation involves a far broader array of communities than does TREC, and thus mechanisms for community-building are even more essential.

A specific proposal for DL evaluation community building that could duplicate the success of TREC, and perhaps overcome some of its limitations, is the following:

The community building initiative funding should develop a powerful DL Evaluation Portal and provide seed support to two or three institutions to engage in the necessary community building efforts. These efforts include organizing regular meetings (these could be workshops at the JCDL) and keeping the portal content fresh (it should be a hub to share testbeds, metrics, evaluation case studies, DL evaluation instruments, success/failure stories, etc.). The latter thrust overlaps with the recommendation on development of DL evaluation infrastructure, while also serving the function of creating a community. The funding can be treated as seed funding with the idea that the institutions will continue to fund the efforts beyond the grant period. Many research universities now have DL programs and staff dedicated to such efforts; the DL staff in the funded institutions could be trained as "community support staff" to take over the responsibility beyond the grant period.

So far the support of NSF for DL evaluation has been helpful to explore and establish the needs as described in this report. Support for community building can help to meet the needs in a concrete and long-term way.

Summary of Recommendations

Research, planning, and deployment of digital libraries all can benefit from evaluation – whether formative, summative, iterative, or comparative. Evaluation efforts can have substantial benefits to digital library development by focusing designers on measurable goals, by providing data on which to reassess those goals, and by assessing outcomes. While many funding efforts have requested or required evaluation, all too rarely is the evaluation actually accomplished. Among the primary reasons for not evaluating information systems is the lack of expertise, the lack of readily available metrics and test beds, and the lack of comparative data on uses, users, and usability. Perhaps most importantly, the nascent community for DL evaluation needs to be nurtured and developed. It is the hope of the workshop participants that future funding initiatives in digital library evaluation will lead to the reduction of these barriers, to a wide array of new measures, metrics, test beds, and substantial understanding of digital library systems and services, and to a community of research and practice that can address the goals of digital library evaluation.

REFERENCES

Berland, G. K., Elliott, M. N., Morale, L. S., Algazy, J. I., Kravitz, R. L.; Broder, M. S.; Kanouse, D. E.; Munoz, J. A.; Puyol, J. A.; Lara, M.; Watkins, K. E.; Yang, H.; & McGlynn, E. A. (2001). Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association*, 285, 2612-2621.

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

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

Bishop, A.P.; Van House, N.; & Buttenfield, B.P.; (Eds.) (In press). *Digital Library Use: Social Practice in Design and Evaluation*. Cambridge, MA: The MIT Press.

Bishop, A. P., Neumann, L. J., Star, S. L., Merkel, C., Ignacio, E., & Sandusky, R. J. (2000). Digital libraries: Situating use in changing information infrastructure. Journal of the American Society for Information Science, 51, 394-413.

Bishop, A. P., Mehra, B., Bazzell, I., & Smith, C. (2000). Socially grounded user studies in digital library development. First Monday, 5 (6), http://www.firstmonday.dk/issues/issue5_6/bishop/index.html

Bishop, A. P., Bazzell, I., Mehra, B., & Smith, C. (2001). Afya: Social and digital technologies that reach across the digital divide, First Monday, 6(4). http://www.firstmonday.dk/issues/issue6_4/bishop/index.html

Borgman, C.L.; Bates, M.J.; Cloonan, M.V.; Efthimiadis, E.N.; Gilliland-Swetland, A.; Kafai, Y.; Leazer, G.L.; Maddox, A. (1996). *Social Aspects Of Digital Libraries*. Final Report to the National Science Foundation; Computer, Information Science, and Engineering Directorate; Division of Information, Robotics, and Intelligent Systems; Information Technology and Organizations Program. Award number 95-28808. Available at: http://is.gseis.ucla.edu/DL/.

Borgman, C.L., Gilliland-Swetland, A.J., Leazer, G.H., Mayer, R., Gwynn, D., Gazan, R., Mautone, P. (2000). Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype (ADEPT). *Library Trends*, 49, 228-250.

Borgman, C.L. Leazer, G.H., Gilliland-Swetland, A.J., & Gazan, R. (2001). Iterative Design and Evaluation of a Geographic Digital Library for University Students: A Case Study of the Alexandria Digital Earth Prototype (ADEPT). In P. Constantopoulos and I.T. Sølvberg (eds.): *Proceedings of the European Conference on Digital Libraries*, Darmstadt, Germany, 5-7 September, 2001. Lecture Notes in Computer Science 2163, Springer-Verlag.

Burstein, L.; & Freeman, H.E. (1985). Perspectives on data collection in evaluations. In L. Burstein, H. E. Freeman, & P.H. Rossi (eds.). *Collecting Evaluation Data.* Beverly Hills: Sage. Pp. 15-34.

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

Computer Science and Telecommunications Board; Commission on Physical Sciences, Mathematics, and Applications; National Research Council. (1997). *More than Screen Deep: Toward Every-Citizen Interfaces to the Nation's Information Infrastructure.* Washington, D.C.: National Academy Press.

Downie, J.S. (2002). Workshop on the creation of standardized test collections, tasks, and metrics for music information retrieval (MIR) and music digital library (MDL) evaluation, 18 July 2002. ACM/IEEE Joint Conference on Digital Libraries, Portland, Oregon.

Druin, A., et al. (2001). Designing a digital library for children: An intergenerational partnership. In Fox, E.A.; & Borgman, C.L. (eds.). *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries.* June 24-28, 2001, Roanoke, VA. New York: ACM. Pp. 398-405.

Fox, E.A.; & Borgman, C.L. (eds.). (2001*). Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. June 24-28, 2001, Roanoke, VA. New York: ACM.

Freire, P. (2002). *Pedagogy of the Oppressed*. 30th anniversary ed. New York: Continuum.

Fuhr, N.; Gövert, N.; Kazai, G.; Lalmas, M. (2002). INEX: Initiative for the Evaluation of XML Retrieval. http://ls6-www.informatik.uni-dortmund.de/ir/publications/2002/Fuhr_etal:02a.html In: Proceedings ACM SIGIR 2002 Workshop on XML and Information Retrieval.

Fuhr, N.; Hansen, P.; Mabe, M.; Micsik, A.; & Sølvberg, T. (2001*). Digital Libraries: A Generic Classification and Evaluation Scheme*, in *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries,* ECDL 2001, Springer LNCS 2163, pp. 187-199.

Gilliland-Swetland, A. J. (1998). Evaluation design for large-scale, collaborative online archives: interim report of the Online Archive of California Evaluation Project. *Archives & Museum Informatics* 12 (3-4): 177-203.

Gilliland-Swetland, A.J., Leazer, G.H. (2001). Iscapes: Digital Library Environments to Promote Scientific Thinking by Undergraduates in Geography. In Fox, E.A.; & Borgman, C.L. (eds*.). Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. June 24-28, 2001, Roanoke, VA. New York: ACM. Pp. 120-121.

Harris, T. M.; & Weiner, D. (1996). GIS and Society: The Social
Implications of How People, Space, and Environment are Represented in GIS. Technical Report 96(7). Santa Barbara, CA: National Center for Geographic Information and Analysis.

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.

IEEE Standard Glossary of Software Engineering Technology. (1991). IEEE Software Engineering Standards Collection.

Kafai, Y.; & Gilliland-Swetland, A.J. (2001). Integrating historical source materials into elementary science classroom activities, *Science Education*, 85, 349-367.

Kirk, J.; & Miller, M.L. (1986). *Reliability and validity in qualitative research*. Newbury Park, CA: Sage.

Landauer, T.K. (1995). *The Trouble with Computers: Usefulness, Usability and Productivity*. Cambridge, MA: MIT Press.

Leazer, G.L., Gilliland-Swetland, A.J., Borgman, C.L. (2000). Evaluating the use of a geographic digital library in undergraduate classrooms: the Alexandria Digital Earth Prototype (ADEPT). *Proceedings of the Fifth ACM Conference on Digital Libraries*, San

Antonio, Texas, June 2-7, 2000. (pp. 248-249). New York: Association for Computing Machinery.

Leazer, G.L., Gilliland-Swetland, A.J., Borgman, C.L., & Mayer, R. (2000). Classroom Evaluation of the Alexandria Digital Earth Prototype (ADEPT). In D.H. Kraft (ed.), *Proceedings of the American Society for Information Science Annual Meeting, 37*, November 12-16, 2000, Chicago. Medford, NJ: Information Today. Pp. 334-340.

Lindland, O.I; Sindre, G; Sølvberg, A. (1994, March). Understanding Quality in Conceptual Modeling. *IEEE Software*.

Marchionini, G.; & Hersh, W. (eds.). (2002). *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. July 14-18, 2002, Portland, OR. New York: ACM.

Morse, E.L. Evaluation methodologies for information management. *D-Lib Magazine*, 8(9), http://www.dlib.org/dlib/september02/morse/09morse.html

Nielsen, J. (1993). *Usability Engineering*. Boston: Academic Press.

Peters, C. (ed.). (2001). Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. *Lecture Notes in Computer Science* 2069, Springer, 387p.

Peters, C.; Braschler, M.; Gonzalo, J.; & Kluck, G. (Eds.). (2002). Second Workshop of the Cross-Language Evaluation Forum. CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised papers. ISSN 0302-9743, ISBN 3-540-44042-9 *Lecture Notes in Computer Science* 2406, Springer, 601p.

Peters, T.A. (ed.) (2000). Assessing Digital Library Services. Special Issue, *Library Trends*, 49, 221-390.

Rogers, E. M. (1986). *Communication technology: The New Media in Society*. New York: Free Press.

Saracevic, T. (2000). Digital library evaluation: Toward evolution of concepts. *Library Trends*, 49, 350-369.

Sawyer, P.; Flanders, A.; & Wixon, D. (1996). Making a difference -- the impact of inspections. *Proceedings of the Conference on Human Factors in Computing Systems*, Association for Computing Machinery. New York: ACM, pp 375-382.

Shim, W., et al. (2001). Measures and statistics for research library networked services: procedures and issues. *ARL E-Metrics Phase II Report*. Washington, DC: Association of Research Libraries.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction,* 3rd ed. Reading, MA: Addison-Wesley.

Sumner, T.; Dawe, M. (2001). Looking at digital library usability from a reuse perspective. In Fox, E.A.; & Borgman, C.L. (eds.). *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries.* June 24-28, 2001, Roanoke, VA. New York: ACM. Pp. 416-425.

Thayer, R.H.; Dorfmann, M. (1990). Glossary. *System and Software Requirements Engineering*, pp605-677. IEEE Computer Society Press.

Whitmore, E. (ed.). (1998). *Understanding and Practicing Participatory Evaluation*. San Francisco, Jossey-Bass.

Williams, F.; Rice, R.E.; & Rogers, E.M. (1988). *Research Methods and the New Media.* New York: The Free Press.

Whyte, W. F. (ed.). (1991). *Participatory Action Research*. Newbury Park, CA: Sage.

All other documents are posted on the workshop web site:

http://www.sztaki.hu/conferences/deval/
–Reports Of Breakout Groups
–Papers Presented At The Workshop
–Call For Papers
–Workshop Agenda
–List Of Attendees