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

Borgman, Christine; Sølvberg, Ingeborg; Kovács, László (Eds.)

Acknowledgements

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.

László Kovács of the Hungarian Academy of Sciences (MTA SZTAKI) was General Chair and host for the workshop. The workshop program was co-chaired by Christine L. Borgman of UCLA, USA, and Ingeborg T. Sølvberg of NTNU, Norway.

The success of this workshop was due to the contributions of many people, especially András Micsik and Zoltán Tóth, who handled the website and the local arrangements.

DELOS workshops

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

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

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.

Table of Contents

I INTRODUCTION

1. Goals of the Workshop

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.

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.

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.

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. Networked information systems are now a 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.

3. Some 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.

Some 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. (http://www.dlib.org/test-suite/)

This short statement suggests that's all the US has done! Please include more here (have pasted in from the NSF report here) and a link to the NSF report.

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.

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).

US. Efforts are discussed further in the report of this workshop to the National Science Foundation (Borgman, 2002).

Borgman, C.L. (2002). Final report to the National Science Foundation. 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. Grant IIS-0225626. http://www.sztaki.hu/conferences/deval/presentations/final_report.html.

Some European Activities on Evaluation of Digital Libraries

DELOS Network of Excellence (DELOS NoE) (www.delos-noe.org) aims at providing a *Digital Library Evaluation Forum* and *Digital Library Test Suites* as one of its main activities. DELOS is an activity within the European Research Consortium for Informatics and Mathematics (ERCIM) (http://www.ercim.org)

In "DELOS Working Group 2: Evaluation" 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. (http://clef.iei.pi.cnr.it/ , http://www.sztaki.hu/delos_wg21/ , http://qmir.dcs.qmw.ac.uk/INEX/)

4. Summary of the 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 (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 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

Two broad themes emerged from the papers, breakout groups, and the rich plenary discussions that took place at the workshop: the complementary needs for

- metrics and test beds
- evaluation in the context of specific digital library applications. `

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.

II PAPERS

The DLib Test Suite and Metrics Working Group: Harvesting the Experience from the Digital Library Initiative

Ronald L. Larsen[1]
University of Maryland

DARPA's DLib Test Suite project[i] was an early attempt at organizing a rigorous and well-supported testbed to enable comparative evaluation of digital library technologies and capabilities. The test suite included a diverse and heterogeneous set of resources deliberately selected to foster research in interoperability, and sponsored a Metrics Working Group (MWG)[ii] to develop quantitative performance measures. The test suite went largely underutilized by the research community, and the MWG, while making significant progress, found the stated objective daunting. Clearly, much remains to be done in both the conception of effective test beds and the instrumentation to assess progress.

The DELOS Workshop on Evaluation of Digital Libraries provides an opportunity to make further progress in this important area, engaging an international community and building on the collective experience accumulated from a larger and more diverse set of digital library projects. In this paper, the DLib Test Suite is briefly reviewed and the progress of the MWG is described.

The DLib Test Suite[iii]

The DLib test suite was conceived to address three needs: (1) lowering the barriers of entry for digital library researchers requiring access to large collections and information management services, (2) providing standard sets of data for quantitative and comparative research, and (3) supporting a distributed environment of heterogeneous resources organized to support interoperability experiments. The Test Suite was most successful on the third objective (supported infrastructure) and partially successful on the second one (standard reference sets). While this was expected to lead to a growing set of researchers interested in evaluation of digital libraries (the first objective), there is scant evidence that this occurred.

Six individual digital library projects participated in the test suite[iv]:

- Carnegie Mellon University's Informedia Digital Video and Spoken Language Document Testbed (digitized and cataloged televised news)
- Cornell University's Networked Computer Science Technical Reference Library (computer science technical reports in a globally distributed set of repositories)
- UC Berkeley's Environmental Digital Library (images, databases, and scanned documents)
- UC Santa Barbara's Alexandria Digital Library (maps, images, and geo-located documents)

[1] The author will be at the University of Pittsburgh, effective July 1, 2002.

- The University of Illinois at Urbana Champaign's Desktop Link to Engineering Resources - DeLIver (online access to scholarly publishers' journals)
- The University of Tennessee – Knoxville's Netlib and the National High-performance Software Exchange (software, numerical databases, and accompanying documentation)

Together, these provided a very diverse set of information resources, services, and interfaces… an environment suitable for creative exploration of interoperability issues.

The model for interoperability experiments among the test suite participants was based on the view of digital libraries as *repositories* of *digital objects*. A *digital object* has a unique and persistent identifier, key metadata describing it, a data stream that can be invoked as a typed sequence of bytes, and a disseminator to map the data stream into a particular form for delivery. A *repository* instantiates digital objects and supports their use in a network environment. It also implements a level of abstraction over the underlying storage mechanisms and provides a secure environment for the management and use of digital objects. Fundamental to the operation of the repositories was a common *Repository Access Protocol (RAP)* that guarantees the integrity of digital objects and facilitates interoperability among repositories.

The DLib Metrics Working Group

The principal focus of the DLib MWG was on information discovery with a human in the loop, in which the information sought is distributed among a heterogeneous set of sources. The objective was to define a set of scientifically rigorous metrics and measures that would enable comparative evaluation of information discovery techniques and algorithms that yielded repeatable results over multiple experiments. As stated by Bill Arms[v], "It should be possible for other researchers to repeat experiments, with different data and different implementations, and to replicate the basic results. The result should be evaluated against relevant, repeatable criteria, so that strengths and weaknesses of alternative approaches can be compared and improvements measured."

The MWG was chartered to consider evaluation issues in the system, user, and content domains. At the systems level, interest focused on interoperability, scalability, heterogeneity, reliability, and integration. At the user level, issues of relevance, specificity, timeliness, effort vs. effect, and usability dominated. In areas of content, measures of sufficiency, currency, and quality were sought.

Scenario-based evaluation was anticipated, and much of the work of the group ultimately revolved around definitions of canonical scenarios. A scenario was defined to include abstract classes, specific instances of those classes, and a common method of scoring. The use of simulation models as well as measurement of real systems was envisioned.

Three sub-groups were identified, although only two ultimately convened. The first sub-group, on metadata issues, focused specifically on metadata for interoperability or sharability, recognizing a spectrum of interoperability issues. Interoperability among

systems using a common standard is clearly the easiest, but rarely fully achievable. More realistic is the expectation of a base standard with extensions to accommodate the specific characteristics of a particular collection or system. The most difficult is clearly interoperability among systems supported by fully divergent metadata sets. Metadata interoperability is required to support: (1) search and retrieval, (2) intellectual property rights management, (3) administration and preservation, and (4) evaluation and use. A system's ability to support interoperability in these areas is fundamentally dependent on the quality of the metadata, and the sub-group explicitly dealt with a range of metadata quality issues, including: (1) specificity, (2) completeness of fields, (3) syntactic correctness, (4) semantic correctness, and (5) consistency, as implemented through authority control.

The second sub-group addressed user-level issues and documented their progress in a series of reports, including (vi) and (vii), summarized below. While the initial charge to the sub-group sought scenario-independent metrics, analogous to precision and recall for information retrieval, such metrics proved to be beyond reach. Scenario-based metrics, while less general, appear to be the best we can achieve at the current state of technology.

The third sub-group was intended to capture the interests of publishers more directly, but in the time the MWG had to conduct its work, the publishers' sub-group was unable to assemble. While the other two sub-groups made efforts to address issues of the publishing community, the development of effective metrics and test beds would benefit from their continued engagement.

Scenario-based Challenge Problems

The MWG solicited challenge problems that would provide measurable scenarios for digital library evaluation. Six categories of challenge problems were sought:

1) Discovery (query formulation and information retrieval)
2) Dissemination (the use of digital libraries for distribution or publication)
3) Other user activities not covered by discovery or dissemination
4) Library administration (scenarios typifying librarians' interactions)
5) System operation (scenarios characteristic of developers or support staff)
6) Other scenarios outside the scope of any of the above.

Scenario definitions included a simple (one sentence) problem statement, a problem description, the issues identified for explicit measurement, and the relevant metrics suggested. Eighteen scenarios were proposed.[viii] An example scenario is "artifact assembly," in which a complex discovery process is described that requires the assembly of a number of information artifacts randomly distributed among sites accessible over the Internet. The artifacts to be acquired have a defined relationship to each other, and all must be retrieved and correlated in order to complete the task. The issues relevant to the scenario include metadata definition and information characterization to allow discovery of the artifacts and recognition of their relationships. The key element behind the challenge problem was the discovery of the *correlated information objects*. The metrics suggested

include precision and recall of the retrieved items, and the elapsed time to complete the query, discovery and correlation tasks.

Other scenarios included:

1) Discovery of reliable, accurate information from sources unknown to the user (e.g., what is the temperature in Troy, NY?).
2) Formulation of a query that, in turn, depends on another set of correlated data, but for which the dependence is data-dependent, rather than source-dependent (e.g., analyzing trends in Far East financial markets, where there exists a correlation between Tokyo and Hong Kong market indicators).
3) Publishing (or disseminating) to an audience with a known characteristic but unknown membership.
4) Querying across media and linguistic boundaries.

Nearly half of the scenarios proposed addressed issues related to discovery. Three scenarios addressed dissemination issues, and five more addressed usage-oriented scenarios. There was only one scenario proposed for library administration. It focused on measuring the usage of electronic articles from specific sources as a function of resource cost. No scenarios were proposed relating to system operations.

DL Metrics

The range of potential metrics relating to digital libraries is immense. In the process of focusing their efforts, the MWG identified at least seven dimensions against which performance could be measured[ix]. These included: (1) system-wide vs. individual services, (2) user interaction vs. underlying system operation, (3) effort vs. effect (net benefit), (4) snapshot vs. session (temporal granularity), (5) capability vs. utility, (6) single user vs. scalability, and (7) collections and content vs. system capability and utility.

The MWG defined a framework for evaluation that addressed the two fundamental phases of an information-seeking scenario, *query* and *retrieval.* For each of these phases, four factors were considered: (1) timeliness, (2) sufficiency, (3) correctness, and (4) effort. The objective became one of identifying indicators for each of these factors that incorporate appropriate (and measurable) metrics. *Timeliness* clearly focuses on speed, and considers both objective measures (e.g., actual elapsed time to complete an operation) and subjective measures (e.g., the user's perception of how long it takes to complete an operation). *Sufficiency* measures the adequacy of the system's response to queries. Recall is the best-known objective measure, but it is typically applied to well-defined finite test collections. Scaling its use up to distributed heterogeneous digital libraries, even through a supported test suite, represents a significantly larger level of effort than was required for the evaluations conducted under the Text Retrieval Conferences (TREC)[x], for example, and yet these have been relatively costly affairs. Alternative means of building instrumented test collections may be required, or different measures entirely may be needed. Sufficiency also has its subjective element. Did the information seeker view the system's responses as adequately comprehensive for the intended purpose? As sufficiency is to recall, *correctness* is to precision. Correctness is intended to gauge the percentage of returned digital objects that actually are appropriate to the query. Subjectively, one asks

the user if the returned objects are right, credible, useful, or reliable. Finally, *effort* addresses the amount of work required by the user to interact with the system, frame the appropriate query, and acquire the objects desired. Objective metrics could address search complexity, including the number of times the user must interact with the system or iterate the query to get it "correct." Subjective measures would consider the user's perception of the level of effort required. Is the system perceived as "easy" to use, for example, or does the user leave in despair, finding the system's operations to obscure to comprehend?

Query-phase Metrics

Each of the four factors (timeliness, sufficiency, correctness, and effort) has a number of components particular to an operation's phase. The MWG found the query phase to pose a rich set of questions with opportunities for metrics. Considering timeliness, for example, leads to measuring the time required to prepare an adequate query, the time for the system to respond to the query, the perceived responsiveness of the interface mediating between the user and the system, the currency of responses (are they current, or up-to-date, as judged by an informed observer?), and the novelty of the responses (would an informed observer recognize them as new or particularly relevant?).

For sufficiency, a measure identified as *availability* is the proportion of sources that the digital library has direct or indirect access to that an informed user would judge as relevant to a particular query. *Interface guidance* addresses the degree to which the system offers useful guidance or options for alternative query formulation. *Coverage* refers to the breadth of system resources that contribute to building the set of returned references. *Actual* and *perceived recall*, as discussed earlier, measure the comprehensiveness of the set of returned items, as measured against a standard, and as perceived by the user.

For correctness, in addition to the traditional *precision* measure (both perceived and actual), the MWG included *interface power*, by which was meant the ability of the user interface to suggest more powerful and correct search terms, strategies, or tactics. A measure of this could be the proportion of suggestions that are actually chosen by the user and that substantially contribute to the resulting set of appropriate responses. Another metric considered was *redundancy*, which measures the proportion of responses that duplicate other material in the same set. As with precision and recall, redundancy can be measured objectively and subjectively (did the user notice actual redundancy or perceive redundancy that was not present?).

Considering effort, potential measures include *interface usability*, *query complexity*, and *response complexity*. Interface usability assesses both objective and subjective measures of a user's ability to efficiently and effectively construct and submit an accurate query. Objectively, one can count the number of queries constructed in the process of finding the sought material or measure the time taken to complete the search. Subjectively, one can ask the user to rate the relative difficulty of using a particular interface. Query complexity is intended to assess the difficulty of formulating the appropriate query for a particularly abstract or complex problem specification, including, for example, the number of search terms required, or the number of iterations required to formulate the successful query. Response complexity attempts to measure the difficulty the user has in interpreting the

returned query response, either by subjective evaluation, or by measuring the time required for the user to take the next step.

Retrieval-phase Metrics

Retrieval is taken to mean the delivery of disseminations of requested digital objects identified as a result of performing a query. The same four overall measures are suggested. For timeliness, the *dissemination time* (the time between the user requesting a dissemination and its presentation to the user) would be measured. For sufficiency, one could consider a metric such as *presentation appropriateness*, where attempts are made to disseminate retrieved objects in a form tailored to a particular audience. This could be as simple as recognizing that a .pdf document will be more useful than a .txt one, or as complex as translating a document into a different natural language. Another sufficiency metric for retrieval is simply *retrievability*. How many of the references returned refer to actual retrievable items?

The correctness metric is *retrieval correctness*, and is defined as the probability that a retrieved dissemination is, in actuality, the correct one. Effort is measured as *selection effort*, or the difficulty the user encounters in selecting or extracting desired disseminations of digital objects from the set of references returned from a query.

Summary

The DLib Metrics Working Group summarized its evaluation metrics for distributed digital libraries as shown in the following table. Are all of these necessary? Together are they sufficient? Can an alternate set of metrics be conceived that are not scenario-dependent?

Evaluation dimensions	Query	Retrieval
Timeliness	Query preparation time Query response time Interface responsiveness Currency Novelty	Dissemination time
Sufficiency	Availability Interface guidance Coverage Actual recall Perceived recall	Presentation appropriateness Retrievability
Correctness	Response correctness Interface power Actual precision Perceived precision Redundancy	Retrieval correctness
Effort	Interface usability Query complexity Response complexity	Selection effort

Table 1 Evaluation Metrics for Distributed Digital Libraries

Much work remains to be done to realize effective measures and metrics in viable digital library test suites and to develop valid and understandable comparisons useful to the digital library community. While the MWG did examine metrics and did suggest scenarios appropriate for some of those metrics, they were not validated in a broader forum, and, perhaps more importantly, *research methods were not addressed.* A spectrum of evaluation approaches can be envisioned, including user surveys, controlled experiments, instrumented systems, and case studies. Fundamental to approaching any rigorous evaluation, however, is consideration of the costs and returns anticipated. Experience from prior evaluation projects (e.g., TREC, SUMMAC[xi] and MUC[xii]) clearly indicates the high costs typically encountered (for collection development, annotation, and administration) in controlled experiments.

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.

---

[i] See http://www.darpa.mil/ito/psum1999/G834-0.html for the DARPA Project Summary
[ii] See http://www.dlib.org/metrics/public/index.html for the record of the MWG's work
[iii] See http://www.dlib.org/test-suite/ for more information about the DLib Test Suite
[iv] The Corporation for National Research Initiatives (http://www.cnri.reston.va.us/) managed the test suite and supported researchers using it.
[v] Arms, William Y., "Replication of Results and the Need for Test Suites," (http://www.dlib.org/metrics/public/PositionPapers/arms.html), January 1998.
[vi] Leiner, Barry M., Leah Lievrouw, Tassos Nakassis, and Mike Sullivan, "Seeker Scenarios for Distributed Digital Libraries," (http://www.dlib.org/metrics/private/papers/Seeker_Scenarios-rev4.htm), October 1999
[vii] DLib Working Group on Digital Library Metrics,"Digital Library Challenge Problems and Metrics" (http://dlib.org/metrics/private/papers/Challenges.html), September 1998.
[viii] See http://www.dlib.org/metrics/private/forms/challenge/c_index.html for complete details of the scenarios proposed.
[ix] Leiner, Barry M., "Types of Digital Library Metrics," (http://dlib.org/metrics/private/papers/types.html), January 1998.
[x] See http://trec.nist.gov/ for complete information regarding the Text REtrieval Conference (TREC)
[xi] See http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/ for the proceedings of the Tipster Text Summarization Evaluation Conference, May 1998.
[xii] The proceedings of the Message Understanding Conference, April 1998, are available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

Digital Library Classification and Evaluation: A Publisher's View Of the Work of the DELOS Evaluation Forum

Michael Mabe
Director of Academic Relations, Elsevier Science &
Visiting Professor, Dept. Information Science, City University, London, UK

ABSTRACT

The work of the DELOS Network of Excellence Workgroup 2.1 on Digital Library Evaluation has produced a generic scheme for the classification of digital libraries. This scheme allows a wide range of library-types to be compared and can be used to develop metrics for cross-evaluation. The library classification and evaluation is shown to be analogous to aspects of strategic marketing. Specifically, the ideas of market segmentation, gap analysis and market extension are shown to be directly applicable to digital library evaluation. The development of a metaLibrary and the next steps of the Workgroup are also surveyed.

1. INTRODUCTION

Evaluation is an important aspect of any process of improvement. It lies at the heart of the business practice of Total Quality Management and underlies a host of public and private sector efficiency measures. In digital library (dl) research, evaluation is an essential requirement for answering the important questions "what is a good dl?" or "how can we make dl's better?" There is little agreement about the definition of a dl nor how such a dl might be evaluated. It was in this context that the DL Evaluation Forum of the DELOS Network of Excellence began its work. Many commentators (especially Marchionini et al. 01) have stressed a human-centred approach to the design of DLs, and this was at the heart of our thinking about evaluation. Libraries are principally there to serve users, so user expectations and uses will be paramount. As a general rule, the nature of the user and the type of use will predetermine the extent and nature of the data or collection, and this in turn will determine the nature of the technology platform and services. This hierarchical description forms the heart of our generic classification scheme (see Fuhr et al. 01).

2. STRATEGIC MARKETING AND DL EVALUATION

Surprisingly, there is considerable similarity between DL evaluation and that of the annual strategic marketing cycle in a commercial business. The strategic cycle is used with the sole intention of improving the fit of products to market, that is, satisfying customer requirements. The fundamental commandment for marketing a product line is asking, "WHO is buying WHAT and WHY". For the information industries, like publishing and librarianship we can reconstitute this fundamental question as "WHO is *using* WHAT and WHY". Answering each dimension (who, what, why) of this statement for a particular product line creates a clear description of the user, the specific product they are using, and the reason behind their use of it. This process is known as market segmentation, and it is the direct analogue of DL classification.

Once a market has been segmented, appropriate performance measures can be adopted. These enable the evaluation of performance of a product within a given segment. Analysis of the outcomes of product performance in a market segment can then be used leading to improve the product fit to the market or how the product is sold. Of all of these processes, the most important is the initial market segmentation, for if this is wrong, all the other results will be invalid.

2.1 Market Segmentation [McDonald & Keegan 97]

Given its importance to the success of any strategic marketing activity, segmentation of the market should be undertaken with a number of broad principles in mind. Market segments must:

- Group together entities that are highly similar to each other

- Be distinct from other segments within the market
- Be describable in terms of their "use" or "purchase"
- Be reachable through sales etc. channels.

These principles can be seen in operation when the basic marketing commandment of "WHO is using WHAT and WHY" is further broken down.

The "WHO" relates to the customer or user and how they may be described or classified. The most traditional approach is through demographics, such as age, sex or some construct such as "adolescents", "retirees". Another approach is socioeconomic; what type of jobs and what types of incomes does the user/customer have? The British newspaper market is a good example of this approach to customer segmentation, with clear socioeconomic divides between the readers of the broadsheets (*The Times, Guardian, Independent* etc.), the middlebrow tabloids (*The Mail, the Express*) and the downmarket red-top tabloids (*The Sun* etc.). Other criteria which can be used include brand loyalty, degree of use (heavy/light) and lifestyle, which has *inter alia* given rise to the regrettable tags of "yuppy" (Young Urban Professional), "dinky" (Double Income No Kids) and "nimby" (Not In My Backyard).

The "WHAT" can be expanded to "WHAT and HOW" and describes the nature of the product and its penumbra of services and delivery. Traditional categories of segmentation include product, price, place, outlet and services.

The "WHY" relates to the purpose of a use or purchase. Traditionally, this dimension of segmentation has been associated with the Second Commandment of Marketing: "Customers don't buy products they seek to acquire benefits." In other words, a customer X buys product Y in order to Z, where Z is the perceived benefit. Thus, "young men" buy "sports cars" in order to "impress their friends", or "show-off to girls" etc., and these "benefits" affect how the car will be designed and sold to that segment ("young men"). Analogous to benefits in this dimension will also be attitudes, perceptions and preferences. All of these can form part of the further segmentation of the "WHY" dimension.

While segmentation is a powerful tool it does need to be applied with some care. Frequently markets segment along lines much more complex than these examples would suggest. An illustration would be the differentiation between different publishing companies within the overall information pyramid (see Fig. 1). Here, the market of information consumers has been segmented by the level of the information (vertical axis) and the size of the potential readership (the horizontal axis). At the top of the pyramid lies research information, with comparatively few readers. At the bottom, lies the mass market of novels, popular non-fiction, magazines etc. While this allows a general segmentation of publishers to correspond with the level of the information, the example of educational information as a vertical slice demonstrates the oversimplification of such an analysis.

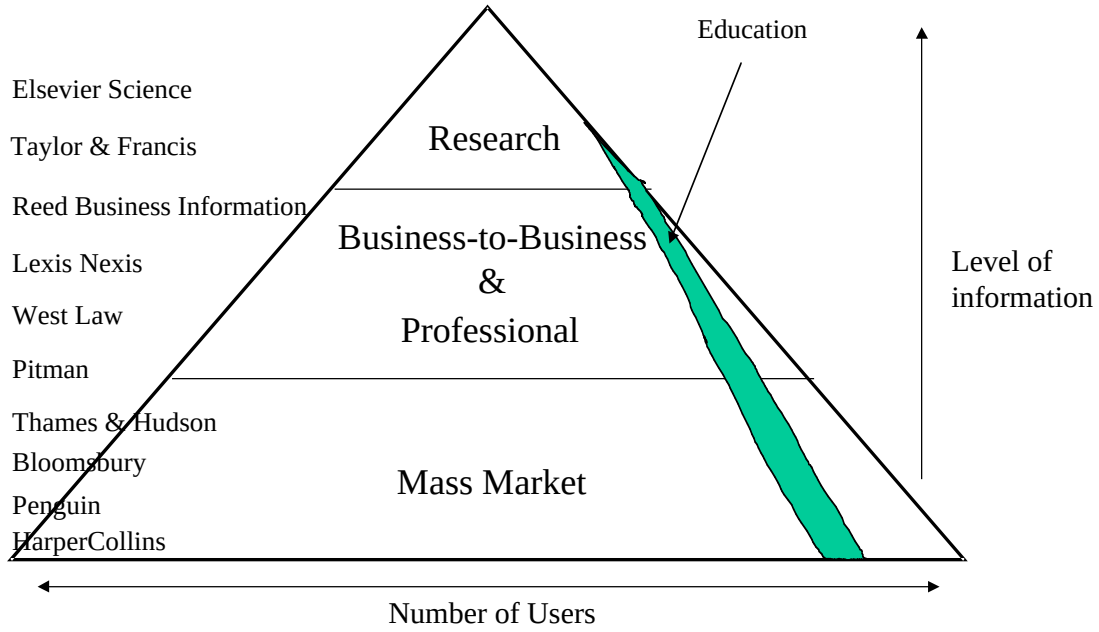


Figure 1. Segmentation and the Information Pyramid

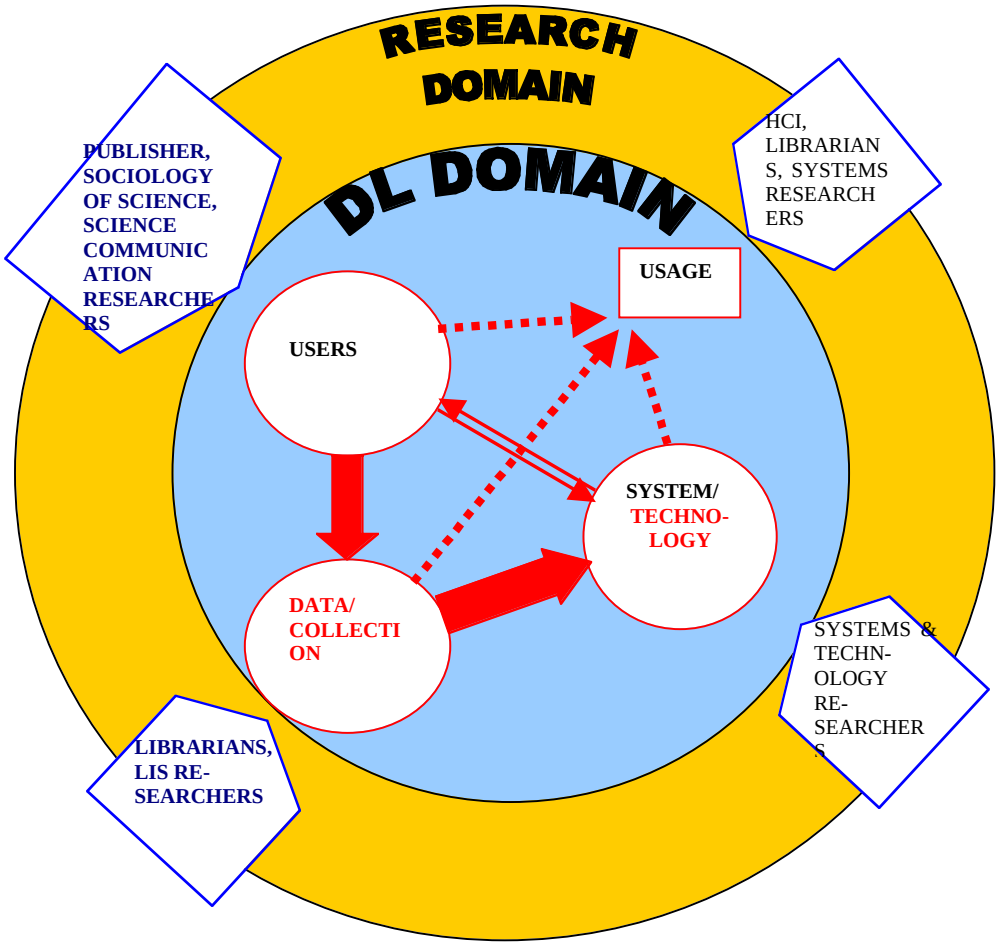


Figure 2. The DELOS NoE Evaluation Forum Classification Scheme

2.2 DL Segmentation/Classification

The DELOS Evaluation Forum took these ideas and applied them to the problem of DL classification. The fundamental commandment of "WHO is using WHAT, HOW and WHY" was a starting point for the first cut of segmentation. For DL purposes, the "WHO" is "users", the "WHAT" is "data/collection", the "HOW" is "technology", and the "WHY" is "uses (purpose)". These primary dimensions could then be further segmented, exactly as in the strategic marketing examples.

- The "users" could be subdivided in terms of *demographics*, their domain of *interest* and the *approaches* that they wished to take towards their information encounter.
- The "data/collection" could be divided into *description* of the collection and the nature of its *management*.
- The "technology" dimension could be classified according to whether it described the technology used (*user technology*, *systems technology*, *document technology*) and *information tools.*
- The "uses (purposes)" relate to the nature of the information *encounter type.* Is the information merely being consumed or used for analysis or synthesis?

Due to the complex nature of the interaction between "users" and "uses", for practical reasons we put these dimensions together as a single element within our classification structure. In the introduction to this paper we referred to the hierarchical relationship between the primary dimensions, with the user being seen as paramount. For DL classification, then, we can rewrite the "Prime Commandment of Marketing" as "WHO and WHY predetermine WHAT and HOW". This leads us naturally to our generalised schema for a DL (see Fig.2 and Fuhr et al. 01).

3. PROPOSED DL CLASSIFICATION AND METRICS

Having come up with a classification scheme we now need to transform it into a mechanism for evaluation. To do this, just as in the strategic marketing example, we have to select relevant measure or metrics to use as the benchmarked qualities. DL metrics are surprisingly similar to a market segment, in that they have to be independent of other metrics, relevant to the dimension being evaluated and above all measurable (compare to distinct, describable in terms of use or purchase, and reachable). The outcome of the work is shown in Table 1.

Table 1. DL Classification Scheme and metrics

Classification	*First dimension*	*Second dimension*	**Metric**
Users/uses	Users	Internal, general, educational, professional, research	Number Distribution
	Domain (subject area)		Distribution
	Information Encounter	Object seeking Browsing	Distribution
	Purpose	Consume Analyse Synthesize	Distribution
Data/Collection	Content	None/partial/full Audio, video, text 2D/3D	Diversity, age, size, quality (white/grey literature)
	Meta content	Bibliographic Indexing/thesaurus/ classification	Media Level of detail
	Management	Rights, workflow user management maintenance	Document age, growth rate, immediacy, completeness, maintenance intervals
Technology	User technology	Document creation, disclosure, interface, browsing, search, printing, group/individual	
	Information access	Retrieval, navigation, filtering, extraction, text mining	Efficiency Effectiveness
	Systems structure technology	Repository Transport models	
	Document technology	Document model Format	

It is noteworthy that three areas of the metrics scheme in Table 1 are blank. These gaps represent the absence any clear methodology for the creation of measurements that relate to these aspects of DL classification and could provide a pointer to new areas of research into measurement methodologies.

4. EVALUATION OF SCHEME AND CREATION OF THE METALIBRARY

The nature of the classification lent itself to the creation of two questionnaires that could be used to determine which DLs are currently available for researchers within the European Union and what test collections would be needed in

the future. The surveys were conducted using a web-based survey software tool, which had been developed by MTA SZTAKI [Kovacs & Micsik 00]. The survey was announced at various mailing lists with major audiences of DL researchers and librarians. It was estimated that around 3-4% responded, of whom about 70% came from the research domain. The surveys showed that while the proposed scheme was appropriate for DL characterization, the wording of the questions in the questionnaire needed further attention. Subsequent surveys have used amended questions to combat this problem.

The results of the surveys have been fed into a database, which is structured according to the generic classification scheme. This so-called MetaLibrary (http://www.sztaki.hu/delos_wg21/metalib/) is an extendable database where classified information about each DL collection or test-bed can be found. The MetaLibrary as a whole provides a snapshot of the current collections as well as indicating where research activity is, and is not, in progress. At the level of individual elements within the classification, the MetaLibrary allows users to determine what collections possess any selected feature. As the MetaLibrary grows in size, it will be a tool to assist DL researchers in finding systems, test-beds and research partners for their needs as well as indicating aspects of DL research that are currently not active.

Once again there is a strong connection with standard strategic marketing practice. The MetaLibrary is functioning as a sophisticated version of the Ansoff Matrix, which is used in gap-analysis to determine future potential market or product extensions.

5. CONCLUSIONS AND NEXT STEPS

There is a clear link between DL evaluation and the techniques of strategic marketing. The tools of market segmentation have been applied to create a novel generalized scheme for describing digital libraries and the research conducted on them. The success of the scheme in practical trials has demonstrated the advantages of a broad and user-centred approach. It has allowed the development of a hierarchy of user-derived metrics and the development of the MetaLibrary of DL collections. This tool could eventually help set the agenda for future DL research, and in doing so contribute to the improvement of DLs for users, the ultimate goal for all forms of DL evaluation.

ACKNOWLEDGEMENTS

The work of the DELOS Network of Excellence is funded by the European Commission.
The members of Work Group 2.1 DL Evaluation Test Suite are:
Norbert Fuhr, U. Darmstadt
Preben Hansen, Swedish Inst. Comp. Sci.
Sarantos Kapidakis, Ionian U.
Laszlo Kovacs, MTA SZTAKI
Michael Mabe, Elsevier Science
Andras Micsik, MTA SZTAKI
Pasquale Savino, IEI-CNR
Ingeborg Srlvberg, Norwegian U. Sci. Techn

REFERENCES

FUHR N, HANSEN P, MABE M, MICSAK A, SOLVBERG I (2001) Digital Libraries: A generic classification and evaluation scheme. In: Constantopoulos P, Solvberg I (eds.) *Research and Advanced Technology for Digital Libraries, 5th Euro. Conf. Digital Libraries 2001, Darmstadt, Proceedings*, Springer, pp.187-199

KOVACS L & MICSIK A (2000) A public service for surveys and decision polls. In: *11th Int. Workshop Database & Expert Systems Applications (DEXA 2000),* pp 307-11

MCDONALD M & KEEGAN W (1997) *Marketing Plans That Work*, Butterworth-Heinemann, Newton, Mass.

MARCHIONINI G, PLAISANT C, KOMLODI A (2001) The people in digital libraries. In: Bishop A, Buttenfield B, VanHouse N (eds.) *Digital Library Use: Social Practice in Design and Evaluation*, MIT Press, Cambridge, Mass. (in press)

Evaluation of Information Access Technologies

Noriko Kando
National Institute of Informatics (NII), Tokyo
kando@nii.ac.jp

Abstract: This paper introduces the activities of *NTCIR* (*NII-NACSIS Test Collections for Information Retrieval and Information Access*) project and the *NTCIR test collections* usable for the systems testing of information access technologies including information retrieval, summarization, question answering. The paper then discusses the design of the evaluation of information access technologies in the aspects of a) a continuum of "*system-oriented*" evaluation and "*user-oriented*" evaluation, and b) context or information seeking tasks in experiments and evaluation. To conclude, some thoughts on the future directions for the test beds of digital libraries research are suggested.

1. Introduction

The *NTCIR Workshop* [1] is a series of evaluation workshops designed to enhance research in information access (IA) technologies including information retrieval (IR), cross-lingual information retrieval (CLIR), information extraction (IE), automatic text summarization, question answering, etc.

The aims of the NTCIR project are;

1. to encourage research in information access technologies by providing large-scale test collections reusable for experiments and common evaluation infrastructures
2. to provide a forum of research groups interested in cross-system comparison and exchanging research ideas in an informal atmosphere, and
3. to investigate methodologies and metrics for evaluation of information access technologies and methods for constructing large-scale reusable test collections.

With the prosperity of the Internet, the importance of research in IA technologies is increasing tremendously. Research and development of IA technologies always require solid evidence based on testing and/or experiments to show the superiority of the proposed system over previous ones. A test collection is a data set used for such testing and experiments.

The importance of large-scale standard test collections in IA research has been widely recognized. Fundamental text processing procedures for IA like stemming and indexing are language-dependent. In particular, processing texts written in Japanese or other East Asian languages such as Chinese are quite different from those in English, French or other European languages since there are no explicit boundaries (i.e., no spaces) between words in a sentence. The NTCIR project therefore started in late 1997 with emphasis on Japanese or other East Asian languages, and its series of workshops has attracted international participation.

Research on digital libraries (DL) covers wide variety of areas including a) information discovery, retrieval and access, b) creation of digital libraries and contents, c) use and users. Evaluation of IA technologies that we have investigated

through the *NTCIR* project is one of the most closely related with DL research and somehow overlapped.

In this paper we introduce the activities of *NTCIR* workshop and its test collections, and then propose some views for the design of the evaluation of IA technologies in the aspects of a) a continuum of "*system-oriented*" evaluation and "*user-oriented*" evaluation, and b) *context* or information seeking tasks in evaluation. The purpose of the paper is to provide a basis of enhanced discussion to construct test beds for digital libraries evaluation by introducing the related example and discussing the views for the evaluation and test beds design.

In the following, the next section provides definition of several terms used in this paper. Section 3 describes the *NTCIR* Workshop and test collections. Section 4 discusses the continuum of the system- to user-oriented evaluation and the context of evaluation design. Section 5 summarizes the discussion.

2. Definition

2.1 Evaluation Workshop

An *evaluation workshop* usually provides a set of data usable for experiments and unified evaluation procedures for experiment results. Each participating group conducts research with its own approach and experiments using the data provided by the workshop organizer. Then the workshop as a whole, wide variety of approaches and techniques are tested on the same data sets by the participating groups. The evaluation workshop also serves as a forum that every participating group learns from each other's experience. One of the most successful examples of evaluation workshop is Text Retrieval Conference (*TREC*) which has been organized by National Institute of Standardization of Technology (NIST), USA since 1991 [2].

The implications of evaluation workshops are various including i) construction of reusable large-scale test collections, ii) facilitating research idea exchange and technology transfer, iii) a "showcase" of the new technologies, iv) motivation of research, v) encouraging intensive discussion and research on evaluation methods, vi) showing the model of experimental design, vii)attracting new comers, and so on [3].

An evaluation workshop is one of the best opportunities to construct large-scale test collections for IR system testing. To evaluate IR systems' search effectiveness, most of the major metrics including recall and precision are calculated based on relevance judgment, and recall can be calculated only when all the relevant documents in the document collection are known. Relevance judgments must be done by human assessors. It is impossible for any assessor to judge relevance to a given search request on every document in a large-scale document collection such as the one containing more than million documents. The different IR systems however can retrieve different relevant documents, and then we can collect candidates for relevant documents to a given search request effectively by pooling the documents retrieved by wide variety IR systems [4], which participated in an evaluation workshop and conducted search of the same search request on the same document collection. Then assessors do relevance judgments on each document in the pool. In this way, the quasi-exhaustive list of relevant documents is created and any relevance-based metrics for search effectiveness including recall and precision can be calculated based on the list. It is also good opportunity for detailed discussion for evaluation methodology and building consensus on evaluation methodology and right answers among the researchers of the domain.

2.2 Information Access

A term "information access (IA)" includes a whole process to make information in the documents usable to the users. A traditional IR system returns a ranked list of retrieved documents which are likely containing relevant information to

the user's information needs. This retrieves relevant documents from a vast document collection and makes these documents usable for the users, and is one of the most fundamental and core process of IA. It is however not the end of the story for the users. After obtaining a ranked list of retrieved documents, the user skims the documents, does relevance judgments, locates the relevant information, reads, analyses, summarizes, compares the contents with other documents, integrates, summarizes and does an information work such as decision making, problem solving, writing, etc., based on the information obtained from the retrieved documents. We have looked such IA technologies to support the users to utilize the information in the large-scale documents in document collections. The scope of the IA is more closely related to the theory and practice of the digital libraries than IR.

3. *NTCIR*

3.1 Brief History of the *NTCIR*

The *NTCIR Workshop* is periodical events which have been taken place once per about one and half years. It was co-sponsored by the Japan Society for Promotion of Science (JSPS) as part of the JSPS "*Research for Future" Program* (JSPS-RFTF 96P00602) and the National Center for Science Information Systems (NACSIS) since 1997. In April 2000, NACSIS was reorganized and changed its name to the National Institute of Informatics (NII). NTCIR was co-sponsored by the JSPS and the Research Center for Information Resources at NII (RCIR/NII,) in FY 2000, and by the RCIR/NII and *Japanese MEXT*[1] *Grant-in-Aid for Scientific Research on Informatics (#13224087)* in and after FY2001.The tasks, test collection constructed, participants of the previous workshops are summarized in Table 1.

For the First NTCIR Workshop, the process started with the distribution of the training data set on 1st November 1998, and ended with the workshop meeting, which was held from 30 August to 1st September 1999 in Tokyo, Japan [5]. The IREX [6], another evaluation workshop of IR and IE (named entities) using Japanese newspaper articles, and NTCIR joined forces in 2000 and have worked together to organize the NTCIR Workshop since then. The challenging tasks of Text Summarization and Question Answering became feasible with this collaboration.

An international collaboration to organize Asian languages IR evaluation was proposed at the 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99). In accordance with the proposal, the Chinese Text Retrieval Tasks are organized by Hsin-Hsi Chen and Kuang-hua Chen, National Taiwan University, at the second workshop, and Cross Language Retrieval of Asian languages at the third workshop.

For the Second Workshop, the process was started from June 2000 and the meeting was held on 7-9 March 2001, NII, Tokyo [7]. The process of the Third NTCIR Workshop starts from August, 2001 and the meeting will be held on 8-10 October 2002, NII, Tokyo[8].

3.2 Focus of the *NTCIR*

Through the series of the NTCIR Workshops, we have looked at both traditional laboratory-typed IR system testing and evaluation of challenging technologies. For the laboratory-typed testing, we have placed emphasis on 1) information retrieval (IR) with Japanese or other Asian languages and 2) cross-lingual information retrieval. For the challenging issues, 3) shift from document retrieval to technologies to utilize "information" in documents, and 4) investigation for evaluation methodologies, including evaluation of automatic text summarization; multi-grade relevance judgments for

[1] MEXT: Ministry of Education, Culture, Sports, Science and Technology

IR; evaluation methods appropriate to the retrieval and processing of a particular document-genre and its usage of the user group and so on.

From the beginning, CLIR is one of the central interests of the *NTCIR*. It was because CLIR between English and own languages are critical for international information transfer in Asian countries, and it was challenging that CLIR between languages with completely different structures and origins such as English and Chinese, or English and Japanese. It was also partly because CLIR techniques are needed even for monolingual text retrieval [9]. For example, a part of a document is sometimes written in English (ex. A Japanese document often contains an English abstract or figure captions, but no Japanese abstract and caption). Technical terms or new terms can be represented in four different forms; i.e., English terms with original spelling, acronyms of the English terms using roman alphabets, transliterated forms of the English terms using Japanese characters, and Japanese terms. The variety in such term expression often causes the decline of the search effectiveness and CLIR techniques are effective to overcome the problem. Moreover, in these years interests towards other Asian cultures has been increased, and importance of the technological information in other Asian countries has been sharply increased in business and industrial sectors.

Table 1. Previous NTCIR Workshops

	period	tasks	subtasks	test collections	particip-ants*		countries
1	Nov.1998-Sept.1999	Ad Hoc IR	J-JE	NTCIR-1	18	28	6
		CLIR	J-E		10		
		Term Extraction	term extraction		9		
			role analysis				
2	June 2000-March 2001	Chinese Text Retrieval	C-C	CIRB010	11	36	8
			E-C				
		Japanese&English IR	monolingual IR: J-J, E-E	NTCIR-1, -2	25		
			CLIR J-E, E-J, J-JE, E-JE				
		Text Summarization	intrinsic - extraction	NTCIR-2Summ	9		
			intrinsic - abstract				
			extrinsic - IR task-based				
3	Aug. 2001-Oct. 2002	CLIR	single lang IR:C-C,K-K,J-J	NTCIR-3CLIR, CIRB020, KEIB010	23	63	9
			bilingual CLIR:x-J,x-C, x-K				
			mulilingual CLIR:x-CJE				
		Patent	cross genre	NTCIR-3Patent	11		
			CLIR CCKE-J				
			optional task				
		Question Answering	task1-basic	NTCIR-3QA	14		
			task2-right anwer				
			task3-serial questions				
		Automatic Text Summarization	single document	NTCIR-3Summ	7		
			multi-document				
		Web Retrieval	survey retrieval	NTCIR-3Web	9		
			target retrieval				
			optional task: output clustering, speech driven				

n-m: n=query language, m=document language(s), J:Japanese, E:English, C:Chinese, K:Korean, x:any
*: number of active participating groups that submitted task resutls

3.3 Test Collections

A test collection is a data set used in system testing or experiments. In the NTCIR project the term "test collections" are used for any kind of data sets usable for system testing and experiments however it often means IR test collections used in search experiments.

The test collections constructed through NTCIR Workshops are listed in Table 2.

Table 2. Test Collections constructed through NTCIR

collection	task	documents			topic		relevance judgment
		genre	size	lang	lang	#	
NTCIR-1	IR	sci. abstract	577MB	JE	J	83	3 grades
CIRB010	IR	nespaper 98-9	210MB	C	CE	50	4 grades
NTCIR-2	IR	sci. abstract	800MB	JE	JE	49	4 grades
NTCIR-2 SUMM	Summ	newspaer94,95,98	180 doc	J	J	-	-
NTCIR-2TAO	Summ	newspaer98	1000 doc	J	J	-	-
KEIB010	IR	newpaper94	74MB	K	CKJE	30	4 grades
CIRB011+020, NTCIR-3CLIR	IR	newspaper98-9	870MB	CJE		50	4 grades
NTCIR-3PAT	IR	patent full'98-9	17GB	J	CKJE	31	3 grades
		+abstract'95-9	4GB	JE			
NTCIR-3 QA	QA	newspaper98-9	282MB	J	J	200+ 800	2 grades
NTCIR-3 SUMM	Summ	newspaper98-9	60 doc	J	J	-	-
NTCIR-3Web	IR	HTML	100GB	J(E)	J	110	5 grades

J:Japanese, E:English, C:Chinese, K:Korean

For example, an IR test collection used in search experiments consists of;

(1) *document collection*

(2) *a set of topics*: written statements of user's search request.

(3) *relevance judgments:* a list of relevant documents for each topic (right answers)

In the retrieval experiments, relevance judgments are most expensive procedure. However, once test collections are created, they can be independent from the settings of the original experiment and can be repeatedly used in the different experiments. Some of the NTCIR test collections contain additional data such as, tagged corpus in NTCIR-1(sentences in the selected documents are segmented manually) and segmented data in NTCIR-2 (every sentence in a whole documents set is automatically segmented into words and phrases beforehand).

3.3.1 Documents

Documents were collected from various domain or genres. Each task carefully selected the appropriate domain of document collection And the task (experiment) design and relevance judgment criteria are set according to each document collection and the supposed user community who use the type of documents in the everyday tasks.

Fig 1 shows an sample record in NTCIR-1 JE. More than half of the documents in the NTCIR-1 JE Collection are English-Japanese paired. Documents are plain text with SGML-like tags in the NTCIR collections. A record may contain document ID, title, a list of author(s), name and date of the conference, abstract, keyword(s) that were assigned by the author(s) of the document, and the name of the host society.

<REC><ACCN>gakkai-0000011144</ACCN>
<TITL TYPE="kanji">電子原稿・電子出版・電子図書館-「SGML 実験誌」の作成実験を通して</TITL>
<TITE TYPE="alpha">Electronic manuscripts, electronic publishing, and electronic library </TITE>
<AUPK TYPE="kanji">根岸 正光</AUPK>
<AUPE TYPE="alpha">Negishi, Masamitsu</AUPE>
<CONF TYPE="kanji">研究発表会(情報学基礎)</CONF>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<CNFD>1991. 11. 19</CNFD>
<ABST TYPE="kanji"><ABST.P>電子出版というキーワードを中心に、文献の執筆、編集、印刷、流通の過程の電子化について、その現状を整理して今後の動向を検討する。とくに、電子出版に関する国際規格である SGML (Standard Generalized Markup Language)に対するわが国での動きに注目し、学術情報センターにおける「SGML 実験誌」およびその全文 CD-ROM 版の作成実験を通じて得られた知見を報告する。また電子図書館について、その諸形態を展望する。出版文化に依拠するこの種の社会システムの場合、技術的な問題というのは、その技術の社会的な受容・浸透の問題であり、この観点から標準化の重要性を論じる。</ABST.P></ABST>
<ABSE TYPE="alpha"><ABSE.P>Current situation on electronic processing in preparation, editing, printing, and distribution of documents is summarized and its future trend is discussed, with focus on the concept: "Electronic publishing: Movements in the country concerning an international standard for electronic publishing. Standard Generalized Markup Language (SGML) is assumed to be important, and the results from an experiment at NACSIS to publish an "SGML Experimental Journal" and to make its full-text CD-ROM version are reported. Various forms of "Electronic Library" are also investigated. The author emphasizes standardization, as technological problems for those social systems based on the cultural settings of publication of the country, are the problems of acceptance and penetration of the technology in the society.</ABSE.P></ABSE>
<KYWD TYPE="kanji">電子出版 // 電子図書館 // 電子原稿 // SGML // 学術情報センター // 全文データベース</KYWD>
<KYWE TYPE="alpha">Electronic publishing // Electronic library // Electronic manuscripts // SGML // NACSIS // Full text databases</KYWE>
<SOCN TYPE="kanji">情報処理学会</SOCN>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>
</REC>

Fig. 1 Sample Document (NTCIR-1, JE)

3.3.2 Topics

A sample topic record used in the CLIR at the NTCIR Workshop 3 is shown in Fig. 2. Topics are defined as statements of "user's requests" rather than "queries", which are the strings actually submitted to the system, since we wish to allow both manual and automatic query construction from the topics.

The topics contain SGML-like tags. A topic consists of the title of the topic, a description (question), a detailed narrative, and a list of concepts and field(s). The title is a very short description of the topic and can be used as a very short query that resembles those often submitted by users of Internet search engines. Each narrative may contain a detailed explanation of the topic, term definitions, background knowledge, the purpose of the search, criteria for judgment of relevance, etc.

<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties of the US National Basketball Association at the end of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include the causes of NBA labor dispute, the relations between the players and the management, main controversial issues of both sides, compromises after negotiation and content of the new agreement, etc. The document will be regarded as irrelevant if it only touched upon the influences of closing the court on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team, league, labor dispute, league and union, negotiation, to sign an agreement, salary, lockout, Stern, Bird Regulation.
</CONC>
</TOPIC>

Fig. 2. A Sample Topic (CLIR at NTCIR WS 3)

3.3.3 Relevance Judgments (Right Answers)

The relevance judgments were conducted using multi-grades. In relevance judgment files contained not only the relevance of each document in the pool, but also contained extracted phrases or passages showing the reason the analyst assessed the document as "relevant". These statements were used to confirm the judgments, and also in the hope of future use in experiments related to extracting answer passages.

In addition, we proposed new measures, *weighted R precision* and *weighted average precision*, for IR system testing with ranked output based on multi-grade relevance judgments [10]. Intuitively, the highly relevant documents are more important for users than the partially relevant, and the documents retrieved in the higher ranks in the ranked list are more important. Therefore, the systems producing search results in which higher relevance documents are in higher ranks in the ranked list, should be rated as better. Based on the review of existing IR system evaluation measures, it was decided that both of the proposed measures be single number, and can be averaged over a number of topics.

Most IR systems and experiments have assumed that the highly relevant items are useful to all users. However, some user-oriented studies have suggested that partially relevant items may be important for specific users and they should not be collapsed into relevant or irrelevant items, but should be analyzed separately [11]. More investigation is required.

3.3.4 Linguistic analysis (additional data)

NTCIR-1 contains a "Tagged Corpus". This contains detailed hand-tagged part-of-speech (POS) tags for 2,000 Japanese documents selected from NTCIR-1. Spelling errors are manually collected. Because of the absence of explicit boundaries between words in Japanese sentences, we set three levels of lexical boundaries (i.e., word boundaries, and strong and weak morpheme boundaries).

In NTCIR-2, the segmented data of the whole J (Japanese document) collection are provided. They are segmented into three levels of lexical boundaries using a commercially available morphological analyzer called HAPPINESS. An analysis of the effect of segmentation is reported in Yoshioka et al. [12].

3.3. 5 Robustness of the System Evaluation using the Test Collections

The test collections NTCIR-1 and -2 have been tested for the following aspects, to enable their use as a reliable tool for IR system testing:

- exhaustiveness of the document pool
- inter-analyst consistency and its effect on system evaluation
- topic-by-topic evaluation.

The results have been reported and published on various occasions [13–16]. In terms of exhaustiveness, pooling the top 100 documents from each run worked well for topics with fewer than 100 relevant documents. For topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents, coverage was higher than 90% if combined with additional interactive searches. Therefore, we conducted additional interactive searches for the topics with more than 50 relevant documents in the first workshop, and those with more than 100 relevant documents in the second workshop.

When the pool size was larger than 2500 for a specific topic, the number of documents collected from each submitted run was reduced to 90 or 80. This was done to keep the pool size practical and manageable for assessors to keep consistency in the pool. Even though the numbers of documents collected in the pool were different for each topic, the number of documents collected from each run is exactly the same for a specific topic.

A strong correlation was found to exist between the system rankings produced using different relevance judgments and different pooling methods, regardless of the inconsistency of the relevance assessments among analysts and regardless of the different pooling methods used [13–15,17]. It served as an additional support to the analysis reported by Voorhees [18].

3.4 NTCIR Workshop 1 (Nov. 1998 -- Sept. 1999)

The first NTCIR Workshop [5] hosted three tasks below;

1. *Ad Hoc Information Retrieval Task*: to investigate the retrieval performance of systems that search a static set of documents using new search topics (J>JE).
2. *Cross-Lingual Information Retrieval Task*: an ad hoc task in which the documents are in English and the topics are in Japanese (J>E).
3. *Automatic Term Recognition and Role Analysis Task*: (1) to extract terms from titles and abstracts of documents, and (2) to identify the terms representing the "object", "method", and "main operation" of the main topic of each document.

In the Ad Hoc Information Retrieval Task, the document collection containing Japanese, English and Japanese-English paired documents is retrieved by Japanese search topics. In Japan, document collections often naturally consist of such a mixture of Japanese and English. Therefore, the Ad Hoc IR Task at the NTCIR Workshop 1 is substantially CLIR, although some of the participating groups discarded the English section and performed the task as a Japanese monolingual IR.

Communications Research Laboratory (Japan)	RMIT & CSIRO (Australia)
Fuji Xerox (Japan)	Tokyo Univ. of Technology (Japan)
Fujitsu Laboratories (Japan)	Toshiba (Japan)
Central Research Laboratory, Hitachi Co. (Japan)	Toyohashi Univ. of Technology (Japan)
JUSTSYSTEM Corp. (Japan)	Univ. of California Berkeley (US)
Kanagawa Univ. (2) (Japan)	Univ. of Lib. and Inf. Science (Tsukuba, Japan),
KAIST/KORTERM (Korea)	Univ. of Maryland (US)
Manchester Metropolitan Univ. (UK)	Univ. of Tokushima (Japan)
Matsushita Electric Industrial (Japan)	Univ. of Tokyo (Japan)
NACSIS (Japan)	Univ. of Tsukuba (Japan)
National Taiwan Univ. (Taiwan ROC)	Yokohama National Univ. (Japan)
NEC (2) (Japan)	Waseda Univ. (Japan).
NTT (Japan)	

Table 3. Active participants for the first NTCIR Workshop

3.5 NTCIR Workshop 2 (June 2000 -- March 2001)

The second workshop [7] also hosted three tasks, and each task was proposed and organized different research group on the topic.

1. *Chinese Text Retrieval Task (CHTR):* including English-Chinese CLIR (ECIR; E>C) and Chinese monolingual IR (CHIR tasks, C>C) using the test collection CIRB010, consisting of newspaper articles

from five newspapers in Taiwan R.O.C.

2. *Japanese-English IR Task (JEIR):* using the test collection of NTCIR-1 and -2, including monolingual retrieval of Japanese and English (J>J, E>E), and CLIR of Japanese and English (J>E, E>J, J>JE, E>JE).
3. *Text Summarization Task (TSC: Text Summarization Arrange):* text summarization of Japanese newspaper articles of various kinds. The NTCIR-2 Summ Collection was used.

Each task had been proposed and organized by a different research group in a relatively independent manner, while maintaining good contact and discussion with the NTCIR Project organizing group, headed by the author. Evaluation, and what should be evaluated, have been thoroughly discussed in a discussion group.

ATT Labs & Duke Univ. (US)	National Institute of Informatics (Japan)
Communications Research Laboratory (Japan),	NTT-CS & NAIST (Japan)
Fuji Xerox (Japan)	OASIS, Aizu Univ. (Japan)
Fujitsu Laboratories (Japan)	Osaka Kyoiku Univ. (Japan)
Fujitsu R&D Center (China PRC)	Queen College-City Univ. of New York (US)
Central Research Laboratory, Hitachi Co. (Japan)	Ricoh Co. (2) (Japan)
Hong Kong Polytechnic (Hong Kong, China PRC)	Surugadai Univ. (Japan)
Institute of Software, Chinese Academy of Sciences	Trans EZ Co. (Taiwan ROC)
(China PRC)	Toyohashi Univ. of Technology (2) (Japan)
Johns Hopkins Univ. (US)	Univ. of California Berkeley (US)
JUSTSYSTEM Corp. (Japan)	Univ. of Cambridge/Toshiba/Microsoft (UK)
Kanagawa Univ. (Japan)	Univ. of Electro-Communications (2) (Japan)
Korea Advanced Institute of Science and Technology	Univ. of Library and Information Science (Japan)
(KAIST/KORTERM) (Korea)	Univ. of Maryland (US)
Matsushita Electric Industrial (Japan)	Univ. of Tokyo (2) (Japan),
National TsinHua Univ. (Taiwan, ROC)	Yokohama National Univ. (Japan)
NEC Media Research Laboratories (Japan)	Waseda Univ. (Japan).

Table 4. Active participants for the second NTCIR Workshop

3.6 NTCIR Workshop 3 (Sept. 2001 -- Oct. 2002)

The third NTCIR Workshop started with the document data distribution in September 2001 and the workshop meeting will be held in October 2002. We selected five areas of research as tasks; (1) Cross-language information retrieval of Asian languages (CLIR), (2) Patent retrieval (PATENT), (3) Question answering (QAC), (4) Automatic text summarization (TSC2), and (5) Web retrieval (WEB). The updated information is available at http://research.nii.ac.jp/ntcir/workshop/.

3.6.1 Cross-Language Retrieval Task (CLIR)

Documents and topics are in four languages (Chinese, Korean, Japanese and English). Fifty topics for the collections of 1998–1999 (Topic98) and 30 topics for the collection of 1994 (Topic94). Both topic sets contain four languages (Chinese, Korean, English and Japanese). Context of the experimental design is "report writing".

Multilingual CLIR (MLIR): Search the document collection of more than one language by one of four languages

of topics, except the Korean documents because of the time range difference (Xtopic98>CEJ).

Bilingual CLIR (BLIR): Search of any two different languages as language and documents, except the searching of English documents (Xtopic98>C, Xtopic94>K, Xtopic98>J).

Single Language IR (SLIR): Monolingual Search of Chinese, Korean, or Japanese. (Ctopic98>C, Ktopic94>K, Jtopic98>J).

When we think of the "layers of CLIR technologies"[19], the CLIR of newspaper articles closely related to the "pragmatic layer (social, cultural convention, etc) " and cultural/social differences among the countries is the issues we should attack in both topic creation and retrieval. For the scientific information transfer, CLIR between English and own language is the one of the biggest interests in East Asian countries. In these years, interests towards social/cultural aspects in East Asia is increasing especially in younger generation. Also technological information transfer among Asia is one of the critical issues in business and industrial sector. According to these changes in the social needs, the CLIR task has changed from English-Japanese Scientific documents to multilingual newspapers and patent documents.

For the next NTCIR Workshop, Korean newspaper articles published in 1998-99 in both English and Korean language will be added, then multilingual CLIR of four languages of Chinese, Korean, English which is published in Asia and Japanese will be feasible.

3.6.2 Patent Retrieval Task (PATENT)

Context of the experimental design is "search for technological trend survey". Regarding "Cross Genre Retrieval", we assumes that someone send a newspaper article clip to a patent intermediary and ask to retrieve the related patents. Search using ordinary topic fields such as <DESC>, <NARR>, etc. are accepted as well as non-mandatory runs. Topic creation and relevance judgments were conducted by professional patent intermediaries who are members of the Information Retrieval Committee at the Japan Intellectual Property Association. The task design was also done with close collaboration with these professionals.

Main Task

- *Cross-language Cross-Genre Retrieval*: retrieve patents in response to newspaper articles associated with technology and commercial products. Thirty query articles with a short description of the search request. Topics are available in Japanese, English, Chinese (simplified, traditional), and Korean.
- *Monolingual Associative Retrieval:* retrieve patents associated with a Japanese patent as input. Thirty query patents with a short description of search requests.

Optional task: Any research reports are invited on patent processing using the above data, including, but not limited to: generating patent maps, paraphrasing claims, aligning claims and examples, summarization for patents, clustering patents.

document:

- Japanese patents: 1998–1999 (ca. 17GB, ca 700K docs)
- JAPIO patent abstracts: 1995–1999 (ca. 1750K docs)
- Patent Abstracts of Japan (English translations for JAPIO patent abstracts): 1995–1999 (ca. 1750K)
-- Newspaper articles (included in topics)

3.6.3 Question Answering Challenge (QAC)

Task 1: System extracts five answers from the documents in some order. One hundred questions. The system is required to return support information for each answer to the questions. We assume the support information is a paragraph, hundred-character passage or document that includes the answer.

Task 2: System extracts only one answer from the documents. One hundred questions. Support information is required.

Task 3: evaluation of a series of questions. The related questions are given for 30 of the questions of Task 2.

3.6.4 Text Summarization Challenge (tsc2)

Task A (single-document summarization): Given the texts to be summarized and summarization lengths, the participants submit summaries for each text in plain text format.

Task B (multi-document summarization): Given a set of texts, the participants produce summaries of it in plain text format. The information, which was used to produce the document set, such as queries, as well as summarization lengths, is given to the participants.

3.6.5 Web Retrieval Task (web)

Survey retrieval is a search for survey and aims to retrieve many relevant documents as possible. *Target retrieval* is a search aiming a few highly relevant documents to get a quick answer for the search request represented as a topic. "Topic retrieval" is a search in response to a search request and "similarity retrieval" is a search by given relevant document(s). In the relevance judgments, one-hop linked documents were also included in the consideration. A topic contain several extra fields specialized to Web retrieval such as a) known relevant documents, b) information on topic author, who is basically relevance assessors.

A. Survey Retrieval (both recall and precision are evaluated)
   A1. Topic Retrieval
   A2. Similarity Retrieval
B. Target Retrieval (precision-oriented)
C. Optional Task
   C1.Search Results Classification
   C2. Speech-Driven Retrieval
   C3. Other

3.6.6 Features of the NTCIR Workshop 3 Tasks

For the next workshop, we planed some new ventures, including:

(1) Multilingual CLIR (CLIR)
(2) Search by Document (Patent, Web)
(3) Passage Retrieval or submit "evidential passages", passages to show the reason the documents are supposed to be relevant (Patent, QA, Web)
(4) Optional Task (Patent, Web)
(5) Multi-grade Relevance Judgments (CLIR, Patent, Web)
(6) Various Relevance Judgments (Web)
(7) Precision-Oriented Evaluation (QA, Web).

(8) Various types of relevance judgments

For (1), it is our first trial of the CLEF [20] model in Asia. We would like to invite any other language groups who wish to join us by providing document data and relevance judgments or by providing query translation. For (3), we suppose that identifying the most relevant passage in the retrieved documents is required when retrieving longer documents such as Web documents or patents. The primary evaluation will be done from the document base, but we will use the submitted passages as secondary information for further analysis.

(4). For patent and Web tasks, we invite any research groups who are interested in the research using the document collection provided in the tasks for any research projects. Those document collections are new to our research community and many interesting characteristics are included. We also expect that this venture will explore the new tasks possible for future workshops.

For (5), we have used multi-grade relevance judgment so far since it is more natural to the users than binary judgments although most of the standard metrics used for search effectiveness are calculated based on the binary relevance judgments. We uses "cumulated gain" [] and proposes new metrics, "weighted average precision" for that purpose. We will continue this line of investigation and will add "top relevant" for the Web task, as well as standard metrics can be calculated by *trec_eval*.

4. Discussion on Evaluation Design

4.1 Continuum of "system-" and "user-"oriented evaluation

"System-oriented" and "user-oriented" are basic categories of evaluation framework. As summarized in Fig. 3, each has own strengths and demerits. Generally speaking system-oriented evaluation is less context dependent, more controllable, repeatable whereas user-oriented evaluation is individual, more context-dependent and including more uncontrollable variables. These two approaches are complementary each other and we need to look both.

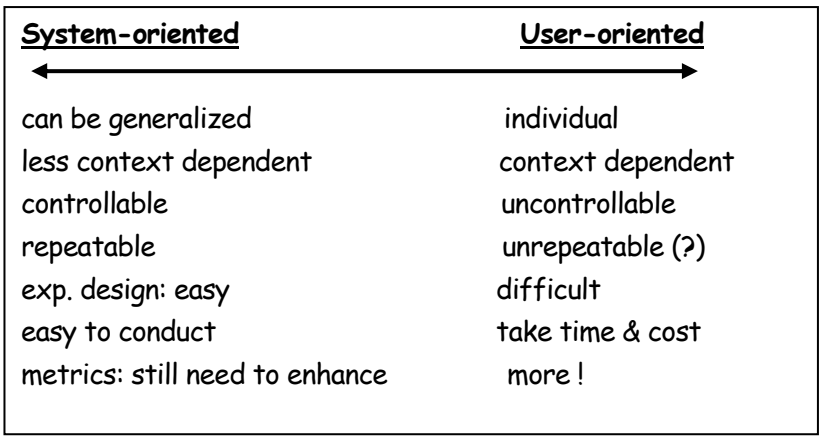


Fig 3. Evaluation of IA technolgoies-1: system-oriented vs user-oriented

These two approaches are not discrepant each other but these two are rather each end of a continuum -- every experiment can be set own position on this continuum between system- and user-oriented according to the purpose of the experiment and available resources. System-oriented evaluation is one of the fundamental approach to test and tune the core functionalities of IA systems. After we train and tune the systems' functionalities based on system-oriented

evaluation repeatedly, we can gradually incorporated user-oriented elements into the experiments and enhance the scope according to each IA system's purpose and design and its situation.

For the user-oriented evaluation, to test every element at a time is almost impossible. Pejtersen [21] proposes step-by-step wise evaluation framework in the context of human-computer interaction. The system designer plans to incorporate some functionalities to the system, for the first predictive evaluation of the new functionality is done repeatedly during the design and define the functionalities and user interface. Then once the functionalities are implemented, assessive evaluation will be done. Through such iteration of the experiment and evaluation, complicated system functionalities can be tested. Through the iterative process of predictive and assessive evaluation of each functionality, the system gradually enhanced its functionality and usability.

For the evaluation of information access systems, we can divide the process into steps then plan the experiments step-by-step. As shown in Fig.4, for example we can start from the evaluation of IR engine, then enhance the scope to post-retrieval processing, then finally pragmatic level of language or document processing such as social or cultural aspects.

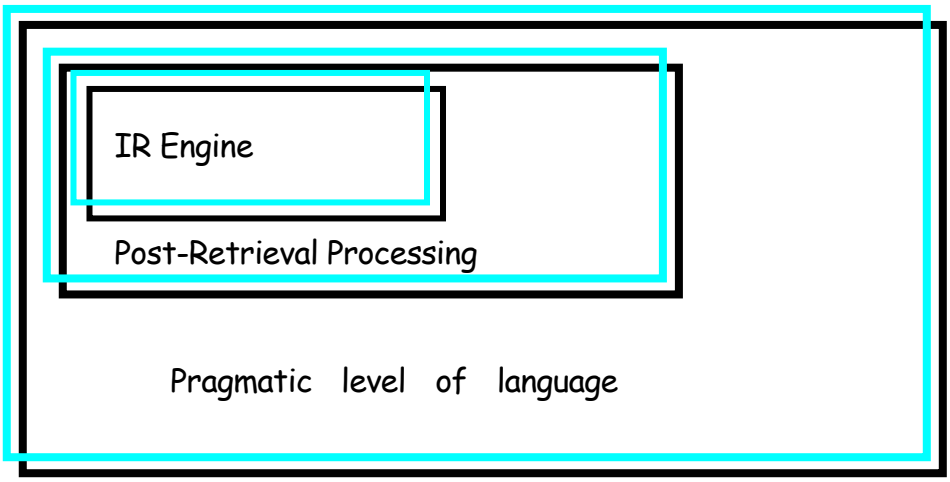


Fig. 4 Evaluation of IA

4.2 Context and task in the evaluation

System-oriented evaluation is "less context dependent" but it is not "context independent". Context, situation, or tasks in the information seeking behavior affect the experimental design and success criteria even in the system-oriented evaluation. IR test collections usable for retrieval experiments are used in laboratory-type system-oriented experiments without real users. However, context, task or situation of the users must be presupposed in the experiments and test collection development.

Each type of documents has own way of usage and characteristics. When we choose a type of documents as a document collection to be retrieved in an experiment, we have to think of what type of users use the type of document with what kind of purpose or tasks, then select appropriate search requests and relevance judgment criteria for that document type in order to make the experiments realistic.

"Relevance" is the most important criteria for the success of retrieval. It must be judged by human assessors even in a

pure laboratory-type testing. It is known that relevance judgments are not stable across the different assessors and can be changed or shifted as time passing even for the same assessor. The criteria of relevance judgments are changed according to each user's situation, task that she/he is involved, purpose or context of the search. There can not be any pure "context independent" relevance judgment and there are no pure "objective" relevance judgments. They must be somehow subjective and done in a particular situation or context. Therefore even for the construction of the test collection usable for laboratory-type testing, test collection and experimental design require to set a context of the search and prepare topics (search request) and relevance judgments appropriate for the type and/or purpose of search, purpose of the system, the types of documents, the usage user community of the document type.

Concept of "relevance" has been discussed and evolved in the history of IR research. Saracevic[22] reviews and discusses the nature of relevance and their definition, then proposes five frameworks to describe the nature of relevance, *i.e.*, 1) systems framework, 2) communication framework, 3) situational framework, 4) psychological framework, and 5) interactive framework. And five categories of relevance have been proposed as manifestations of relevance, *i.e.,* 1) system or algorithmic relevance, 2) topical or subject relevance, 3) cognitive relevance or pertinence, 4) situational relevance or utility, and 5) motivational or affective relevance.

Mizzaro[23] reviews the literature on "relevance" and summarizes the discussion. He proposes four dimensions to construct "relevance", 1) information (I_1:Surrogate, I_2:Document, I_3:Information), 2)user (U_1: Query, U_2: Request, U_3:Information need, U_4: Problem) 3) domain (D_1:Topic, D_2:Task, D_3:Context), and 4) time and each dimensions has one to four elements. The most of definitions of relevance discussed and proposed in the literature can be described and characterized using these dimensions.

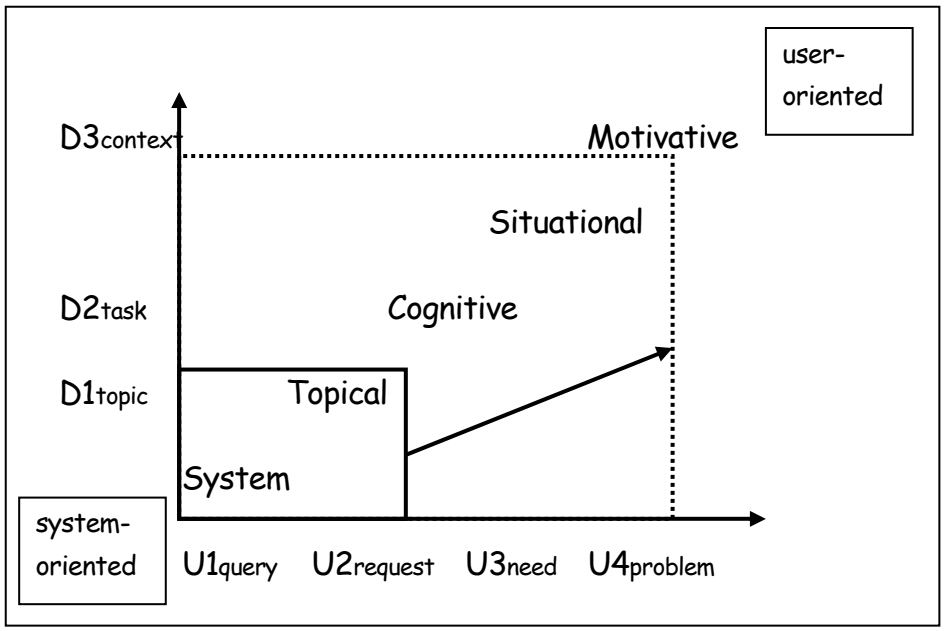


Fig. 5. Scope of relevance: Laboratory-type testing can be extended to situation-conscious

Both Saracevic and Mizzaro propose that "relevance" is not a simple entity but rather complex of several aspects and characteristics. Saracevic's categorization of manifestation of relevance can be related to the Mizzaro's dimensions as follows [24]:

1) System relevance -- I_1:Surrogate , D_1:Topic, U_1: Query
2) Topical relevance -- $I_{1\text{-}2}$:Surrogate & Document , D_1:Topic, U_2: Request
3) Cognitive relevance -- $I_{1\text{-}3}$:Surrogate & Document & Information, $D_{1\text{-}2}$:Topic & Task, U_3:Need
4) Situational relevance -- $I_{1\text{-}3}$:Surrogate & Document & Information, $D_{2\text{-}3}$:Task & Context, U_4: Problem
5) Motivational relevance -- $I_{1\text{-}3}$:Surrogate & Document & Information, D_3:Context, U_4: Problem

The Fig. 5 shows the regions of relevance expressed by combination of Saracevic's and Mizzaro's. Traditionally relevance judgments in IR test collections are done based on somewhere between *system* and *topic relevance*. And if we incorporate the situation, the task that the user involved, or the purpose of the search, etc, we can extend and evolve the scope of the relevance to *situational relevance* even in the static judgments in the test collections although experiments including interaction between users and systems is ineviable to test the IR systems' effectiveness against *motivational relevance.*

5. Future Directions

This paper introduced the NTCIR Workshop and test collections constructed through the workshop for IA technologies evaluation, and discussed the framework for evaluation design in the aspects of a) a continuum of "system-" and "user-oriented" evaluation and b) context or task in the evaluation. The experience of NTCIR and other evaluation workshop or campaign like TREC and CLEF, the implications and contributions of the standard test collections and evaluation workshop are obvious. They encourage the research and development of IA technologies in many ways.

We can expect various effects and implications for the test beds for digital libraries research, and its necessity is obvious. However, the areas and problems covered by the digital library research are wider and more heterogeneous than IR or IA technologies evaluation. Therefore careful planning and designing are inevitable for constructing them. The followings are brief list of the issues to be considered for the digital library test beds. They are just examples of the issues based on the experience of IR and IA test collections and evaluation, and they are not exhaustive.

1. Start from the simplest form of the test bed .
2. Analysis of the stableness and robustness are inevitable when the new metrics and evaluation methodologies are proposed. Their scope and limits must be clearly expressed.
3. Define the issued to be evaluated clearly, -- efficiency, effectiveness, satisfaction, utility, and so on.
4. Context, the user's information seeking task, the user community, content types are all needed for the explicit experiment and test bed design.
5. Consideration on cross-language information access, international collaboration, and cross cultural aspects are needed for digital libraries research.

Regarding the simplicity, and do not incorporate many aspects into one test bed at a time. Evaluate one aspect in an experiment, and then enlarge the scope gradually. Try to make generalizable and robust evaluation suite. The simple test beds can be used in various ways according to the settings that the experimenters set. Since digital libraries are global entities and can be accessed from anywhere in this world. Research and practice of digital libraries also require the point of view of global, cross-lingual, and cross-cultural information transfer. For this aspect, international collaboration in research environment and test beds preparation are worth considered.

References:

[1] NTCIR Project: http://research.nii.ac.jp/ntcir/

[2] TREC. http://trec.nist.gov/

[3] Smeaton, A.F. and Harman, D. "The TREC (IR) experiments and their impact on Europe", *Journal of Information Science*, No. 23, pp 169-174, 1997.

[4] Sparck Jones, K., Rijsbergen, C.J. *Report on the need for and provision of an 'ideal' information retrieval test collection*, Computer laboratory, Univ. Cambridge., 1975 (BLRDD Report)

[5] *NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Tokyo, Japan, 30 Aug.–1 Sept., 1999. ISBN4-924600-77-6. (http://research.nii.ac.jp/ntcir/workshop/ OnlineProceedings/).

[6] IREX URL:http://cs.nyu.edu/cs/projects/proteus/irex/

[7] *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, June 2000–March 2001.ISBN4-924600-96-2. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/)

[8] *NTCIR Workshop 3 Meeting: Working Note of the Third NTCIR Workshop Meeting*, Tokyo, Japan, Oct.8-10, 2002. 6 vols.

[9] Kando, N.: "Cross-linguistic scholarly information transfer and database services in Japan". Presented in the panel on Multilingual Database in the Annual Meeting of the American Society for Information Science, Washington DC. , USA, November 1997.

[10] Kando, N., Kuriyama, K., Yoshioka, M. "Evaluation based on multi-grade relevance judgments". *IPSJ SIG Notes,* Vol.2001-FI-63, pp.105-112, July 2001. (in Japanese w/English abstract)

[11] Spink, A., Greisdorf, H. "Regions and levels: Measuring and mapping users' relevance judgments". *Journal of the American Society for Information Sciences,* Vol.52, No.2, pp.161-173, 2001

[12] Yoshioka, M., Kuriyama, K., Kando, N.: "Analysis on the usage of Japanese segmented texts in the NTCIR Workshop 2." In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000–March 2001 （ISBN：4-924600-96-2).

[13] Kando, N, Nozue, T., Kuriyama, K., Oyama, K., "NTCIR-1: Its policy and practice", *IPSJ SIG Notes*, Vol.99, No.20, pp.33-40, 1999. (in Japanese w/English abstract)

[14] Kuriyama, K., Nozue, T., Kando, N., Oyama, K.: "Pooling for a large scale test collection: Analysis of the search results for the pre-test of the NTCIR-1 Workshop", *IPSJ SIG Notes*, Vol.99-FI-54, pp.25-32 May, 1999 [in Japanese].

[15] Kuriyama, K., Kando, K. "Construction of a large scale test collection: Analysis of the training topics of the NTCIR-1", *IPSJ SIG Notes*, Vol.99-FI-55, pp.41-48, July 1999. (in Japanese w/English abstract)

[16] Kando, N., Eguchi, K., Kuriyama, K., "Construction of a large scale test collection: Analysis of the test topics of the NTCIR-1", In *Proceedings of IPSJ Annual Meeting* [in Japanese]. pp.3-107 -- 3-108, 30 Sept.–3 Oct. 1999.

[17] Kuriyama, K., Yoshioka, M., Kando, N., "Effect of cross-lingual pooling". In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000–March 2001 （ISBN：4-924600-96-2)

[18] Voorhees, E.M., "Variations in relevance judgments and the measurement of retrieval effectiveness", In *Proceedings of 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*., Melbourne, Australia, August 1998, pp.315-323.

[19] Kando, N. "Towards real multilingual information discovery and access ". Presented at ACM Digital Libraries and ACM-SIGIR Joint Workshop on Multilingual Information Discovery and Access. Panel on the Evaluation of the Cross-Language Information Retrieval. Berkeley, CA, USA, August 15, 1999. (http://www.clis2.umd.edu/ conferences/midas/papers/kando2.ppt)

[20] CLEF: Cross-Language Evaluation Forum, http://www.iei.pi.cnr.it/DELOS/CLEF

[21] Rasmussen J., Pejtersen, A.M., Goodstein, L.P., eds. *Cognitive Systems Engineering*. Wiley, 1994.

[22] Saracevic, T. "Relevance reconsidered '96", In *Proceeding of the 2nd Conference on Library and Information Science (CoLIS-2)*, Copenhagen, Denmark, Oct. 1996, p.201-218

[23] 43. Mizzaro, S. "Relevance: The whole history" *Journal of the American Society for Information Science*, Vol.48, No.9, pp.810-32, 1996

[24] Nozue, T., Kando, N. “Primary considerations in the concept of relevance: Relevance judgement of NTCIR”. *IPSJ SIG Notes*, 99-FI-53, Vol.99, No.20, March 1999, p. 49-56. (in Japanese w/English abstract)

A Framework for Criteria and Measures for Evaluation of User Interfaces in Digital Libraries

Nicholas J. Belkin
School of Communication, Information & Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071, USA
nick@belkin.rutgers.edu

Abstract

1 Introduction

Although there has been consistent (if rather low-key) interest in interface design issues in digital libraries (DLs), there has been relatively less attention paid to how such interfaces should be evaluated. Although evaluation of interfaces to DLs does indeed occur, it is only rarely that the question of what criteria and measures are actually appropriate for these very special kinds of systems arises. Rather, either general classes of measures from standard human-computer interaction (HCI) evaluation guidelines, or standard information retrieval (IR) measures are used. In this paper, I propose a framework for relating the DL situation to evaluation of DL interfaces, and make some concrete suggestions, based on previous and ongoing research, as to some criteria and measures which would be appropriate.

2 Background

In the IR and DL literature, it has been the general case that designers (and evaluators) have treated traditional system functionality (e.g. collection, representation, structuring, retrieval, presentation, dissemination of information objects) separately from the interface to the DL or information system. Dumais (1996) and others have argued persuasively that this is a major mistake, and that the interface should be considered an integral part of the system and its functionality, and not merely an appendage or add-on. This argument has been made primarily with respect to the design of systems; here, I wish to make it with respect to their evaluation, as well. I wish also to take seriously the bromide that a system's goal and context should have a strong influence not only on its design, but also on its evaluation. But neither of these statements should be taken to suggest that the only form of evaluation that is relevant to this context is summative, referring to the system as a whole. Rather, they suggest that formative or analytic evaluation, when undertaken, must consider factors relevant to the system's goal as a whole, and to the context in which it exists. Taken together, these positions suggest that a reasonable approach to evaluation of interfaces to DLs is first to identify their goals, and then the features of their context salient both to their overall goals, and to the aspects of the interface which could have influence on achieving those goals.

In general, when evaluation of user interfaces in DLs has been done, it has concentrated on usability as the major criterion, with common HCI-oriented measures being applied. These typically include such "objective" measures as error rate, and "subjective" measures such as perceived ease of use. At times, effectiveness has been used as a criterion, with time to complete a task a typical measure. However, very few such investigations have used measures which relate to the DL goals themselves, in particular with respect to performance of a "real" DL task. The intent of this paper is to suggest how to move from "mere" interface evaluation to evaluation of the user interface in the whole system context.

3 Goals and salient features of the digital library context

Clearly, it is not possible to enumerate goals for DLs in general, since each system will have established its own specific goals, with respect to its user community and its collection(s). This means that the starting point for the evaluation of any aspect of a DL will be the enumeration and specification of its own, particular goals. Having done this, it will then be possible to identify those which can be addressed by the different components and processes of the system. For instance, the goal of exhaustive coverage of the information associated with the DL's domain is unlikely to be impacted by the user interface if the materials are of one type, and to be held in

one place; however, the fact that this is a goal, and therefore a characteristic of the DL, will impact the design, and therefore the evaluation of the user interface to the collection. If the goal is to be achieved by accessing a variety of remote collections, and/or if the materials are of a variety of types, then this goal of exhaustive coverage is indeed one whose achievement can be directly evaluated through the user interface.

Despite the inability to identify general goals for DLs, it is still possible to identify a set of general *criteria* according to which achievement of the goals can be measured. These will include at least efficiency, effectiveness, satisfaction and usability. It seems likely that user interfaces can be evaluated with respect to all of these criteria with respect to a variety of goals.

Although general goals of DLs are not specifiable, it may be possible to identify some general characteristics of the contexts of DLs which are common (or becoming common) to most of them, and which seem especially relevant to the issue of evaluation of the interface in the DL. These include:

- Distributed, heterogeneous (across), yet homogeneous (within) databases
- Focused user communities
- Disparate user types
- Wide variety of task contexts, wide variety of information seeking strategies
- Wide variety of types of information objects (e.g. text, image, video, sound)

These characteristics of DLs place especially severe constraints on their user interfaces, and establish criteria on which they should be evaluated. These criteria interact in complex ways with those associated with the overall goals of the DLs. For instance, that the databases to which the user of a DL will have access are distributed has led to the typical assumption that the results of searches on the different databases should be integrated into a single retrieval display. This has then led to evaluation of such displays, in the user interface, in terms of how well they are integrated, and perhaps how well the result display can be used. The issue of heterogeneity across databases has been dealt with primarily at an algorithmic rather than interface level, and the issue of homogeneity within databases has been largely ignored. But, when all these characteristics are explicitly considered, and when the user interface is then evaluated with respect not only to its usability, but also to the performance of a task which the DL is supposed to help the user to accomplish, the very assumption of the desirability of an integrated result can be brought into question. Indeed, evaluation of the interface with respect to the task, as well as other criteria in Park's (2000) study demonstrated that the homogeneity and heterogeneity characteristics led not only to user preference for interaction with single databases, but also to a performance benefit.

4 A framework for evaluation of user interfaces in digital libraries

The discussion above suggests the following general framework for evaluation of user interfaces in DLs.

1. Identify and specify *all* the goals of the DL in detail. Consider how achievement of these goals can be measured with respect to at least the criteria of efficiency, effectiveness, satisfaction and usability. Consider how each of these goals can be facilitated by the user interface, and consider how each of them can impact on the design and functionality of the user interface.
2. Identify the specific characteristics of this DL in terms of at least databases, community, users, uses, and information objects. Consider how each of these interacts with the goals of the DL, especially in terms of user interface functionality.
3. Specify explicit measures for evaluating the user interface with respect to at least the four criteria of efficiency, effectiveness, satisfaction and usability, such that the measures evaluate achievement of DL goals, as well as interface capabilities and characteristics.

5 An example evaluation

I present an outline of how elements of the framework proposed above were used in an experiment to investigate one aspect of the functionality of interfaces to digital libraries. The focus of this discussion is on how the experimental design and the evaluation measures were arrived at, with only brief mention of the results. This work is reported in much more detail in Park (2000).

The goal of the experiment was to test the hypothesis that users of DLs consisting of several different databases will prefer, and perform better with, an interface which allows them to interact with the databases as if they were

one, to one which requires interaction with each separately. This hypothesis is typically understood to mean that when searching for information in such a DL, the user should need to formulate only one query, which is sent to all of the databases, and that the search results from each of the databases should be integrated and presented as a single results list. The experiment was conducted in a simulated DL, which was constructed especially to test this hypothesis.

The goals of the simulated DL were understood to be to support effective and easy access to a variety of information resources concerned with current events and US government actions and debates. The users of the DL were assumed to be members of the general public interested in such issues, and in their relationships. These assumptions allowed the DL to be constructed using the resources of the TREC test collection (cf. Voorhees & Harman, 1997), which in our case was specifically the following full-text databases:

- Federal Register 1994
- Wall Street Journal 1990-92
- Financial Times 1991-94
- Congressional Record 1993

These were chosen because each has quite different characteristics from the others (reflecting heterogeneity across the databases), each is internally fairly homogeneous in structure and content, and they cover roughly equivalent time periods. In addition, it was possible to identify TREC topics (queries for which relevant and non-relevant documents have already been identified) which had relevant documents distributed in a variety of ways across the databases. InQuery version 3.2 (cf. Callan, Croft & Harding, 1992) was used as the underlying indexing and retrieval system, with a locally developed interface.

In keeping with the goal of the experiment, two general interface/functionality combinations were constructed, each of which could be assumed to influence effectiveness, usability, efficiency and satisfaction. One was an interface through which each of the databases could be accessed separately, but with a common look and feel, and a common interaction structure. This interface allowed the searcher to choose which database to search, and also had a facility for saving and reusing queries. The second interface sent looked just the same, with all the same functionality except that it had no means for selecting a database. Instead, queries were sent to each of the four databases, and the results of the separate searches were cumulated into a single ranked list using the technique described by Xu & Callan (1998). The former interface, by allowing users control over what was to be searched, was thought likely to increase satisfaction. But it was also thought likely to reduce efficiency, usability and effectiveness, because more interaction would be required to achieve desired results.

Within this framework, the experiment was set up in which 28 subjects each searched on six different topics, three using one of the interfaces, and three more using the other. The subjects' task in each search was to enumerate the different *aspects* of the topic, by identifying documents which discussed each of the aspects (this is the so-called aspectual recall task; cf. Over, 1997). Effectiveness in the task was measured by the number of aspects that were correctly identified by the searcher (aspectual recall), by the subjects' satisfaction with the results of each search, and by the subjects' report of which system they found more useful. Satisfaction with the interface was measured by preference for one system over the other in general. Efficiency was measured by the amount of required to do each search, and whether subjects felt they had enough time. Usability was measured by amount of interaction, subjects' perception of ease of search, and which of the systems was easier to use and easier to learn to use.

The measures that were used to evaluate and compare the two interfaces to one another (i.e. to test the initial hypothesis) combined standard types of usability measures from HCI studies, with a somewhat non-standard IR performance measure. These measures were thought to cover the range of issues raised by the hypothesis, and how the interface could affect the user-DL interaction in the specific type of DL that we envisioned.

The results indicated that the general assumption that integrated interaction is needed in DLs of this type should at the very least to be revisited. Performance was in fact better in the common than the integrated interface. But, it was found that users might choose integration if there were better characterization of the databases to be integrated, and that some users prefer control over database selection, while still opting for merging of results. These results demonstrate the value of the framework proposed here in several ways, perhaps most importantly in that they indicate features and characteristics of the situation whose importance was not appreciated at the beginning of the study, but which emerged from the data and their analysis.

6 References

Callan, J.P, Croft, W.B. & Harding, S.M. (1992) The INQUERY retrieval system. In *DEXA 3: Proceedings of the third international conference on database and expert systems applications* (pp. 83-87). Berlin: Springer Verlag.

Dumais, S. T. (1996) Interface Design for Information Retrieval. In *Workshop on Social Aspects of Digital Libraries*, UCLA, February 1996. Retrieved from http://is.gseis.ucla.edu/research/dl/dumais.html on 15 April 2002.

Over, P. (1997) TREC-5 interactive track report. In D. Harman (Ed.) Proceedings of the fifth Text REtrieval Conference (TREC-5) (pp. 29-56). Washington, DC: GPO.

Park, S.-Y. (2000) Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: implications for digital libraries. *Journal of the American Society for Information Science, 51 (5)*, 456-68.

Voorhees, E.M. & Harman, D. (1997) Overview of the fifth text retrieval conference (TREC-5). In D. Harman, (Ed.). *Proceedings of the Fifth Text REtrieval Conference (TREC-5)* (pp. 1-28). Washington, DC: GPO.

Xu, J. & Callan, J.P. (1998) Effective retrieval with distributed collections. In A. Moffat & J. Zobel (Eds.) SIGIR 98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (pp. 112-120). New York: ACM.

EVALUATION METRICS FOR USER-CENTERED RANKING OF CONTENT IN METADLs

James Ford, Fillia Makedon, Li Shen, Tilmann Steinberg
The Dartmouth Experimental Visualization Laboratory (DEVLAB)
Department of Computer Science[1]

Andrew Saykin, Heather Wishart
Brain Imaging Laboratory, Dartmouth Medical School

Dartmouth College
Hanover, NH 03755, USA

I. INTRODUCTION

MetaDLs are digital libraries (DLs) designed to store and manage the access of metadata rather than original data [1]. They are particularly suited for data that is sensitive in nature (possibly proprietary), complex in structure, of high value (or production cost), and highly heterogeneous. The authors describe such a MetaDL system architecture in [9], where the model is demonstrated in the case of human brain data that can include brain activation maps, lesions, brain structures, test methodologies, patient information and tools. This paper describes evaluation metrics for such a MetaDL in computational neuroscience research. MetaDLs are applicable to complex data outside of neuroscience as well: gene sequences in genetic mapping, satellite earth/weather patterns, etc. (Other metadata DL descriptions have been given in [5, 7].)

A MetaDL exists within a two-tier architecture that supports two endeavors: searching for data (and methods) via metadata, and sharing this data in a secure fashion. Tier 1 consists of autonomous DLs containing data, each with an interface allowing it to specify access conditions. Tier 2 systems contain data about the Tier 1 DLs and permit browsing and searching for data that are contained in Tier 1 DLs. MetaDLs use meta-data to describe data objects in a uniform, structured manner, thus deriving a parameterized description of the object. Objects are then accessed, traded, manipulated, updated, etc. securely, as the original data resides with the content owner. Besides protecting the information from malicious abuse, a key additional benefit is making heterogeneous data interoperable through a common description. MetaDLs enable a user to search, browse, manipulate, and rank information, at Tier 2 level, without needing the original. Actual data exchange or access takes place at Tier 1 level, after a "data request" has been formulated and recorded by MetaDLs [1]. MetaDLs are designed to support, document and manage user-to-user transactions on a small scale.

The BrainMap system [2] and the fMRI Data Center [3] attempt to address this situation in the area of functional imaging by consolidating results and data, respectively, in central repositories. BrainMap contains summary information for various refereed publications related to HBM, including for example reported foci of activations in different studies, the number of subjects used in each, and other similar information. Users of the system can use this data for meta-analysis of the results, for instance by grouping several studies to increase statistical power. The system does not provide or facilitate exchange of raw data, however, and this limits the range of possible meta-analyses. It also does not include descriptions of disease-induced structural changes. The fMRI Data Center is an NSF-sponsored site that publishes raw data related to published functional imaging articles in certain refereed journals. These studies are not focused on the Human Brain and are not comprehensive by disease. Authors of published articles are required to provide their data for public access, although no particular format or level of documentation is required. It is conceivable that these data might be used for re-analysis or for combination with data from other studies, but this is hampered by the fact that available data is not indexed by features. Other related projects include METAe [5], MUMIS [6], Dublin Core [7], and ARION [8].

In [1], the authors have outlined a new digital library approach, BrassDL (BRain Access Support System Digital Library) that addresses the limitations of BrainMap and the fMRI Data Center. BrassDL aims to support collaboration, scientific exchange and discovery by a system by representing information via metadata (number of subjects, freshness of data, location, scanner parameters, etc.). Thus, brain structures (lesions, structures) and data, experiments, methods, and subject information are all represented using a common metadata representation that is inserted into the system directly by the researcher. BrassDL provides strong incentives for participation in this pooling of information through a framework of services to the research community. Besides offering visibility, ranking, and feedback, BrassDL can support meta-analysis of metadata information and the BrassDL–Interchange, a business model for managing data interchange and tracking.

In the rest of the paper, section II introduces the BrassDL, a MetaDL of human brain metadata that integrates

[1] This work has been supported in part by NSF-IDM research grant number 0083423.

heterogeneous brain imaging results, methods, experimental conditions, and subject information. Section III introduces a new metric for evaluating the content of a MetaDL in terms of how it matches user demand. Section IV concludes the paper.

II. BRASSDL: A METADL OF HUMAN BRAIN METADATA

As a MetaDL, BrassDL [1] exists within a two-tier architecture. Tier 1 consists of autonomous DLs controlled by data providers (or BrassDL Partners). These DLs contains primary data, metadata and information about their access conditions. BrassDL is a Tier 2 system and contains data about the Tier 1 DLs. BrassDL manages metadata collection, retrieval and evaluation, and also manages the user-to-user data exchange in a smaller scale. BrassDL collects features and parameterizes the evaluation of the dataset, thus enabling a ranking of the results as well as of the user demand, on the basis of how good the system is to the user's request, and according to user-defined preferences.

Efficient data collection is achieved through uniform and structured representation of diverse brain data: activations, lesions, brain structures, methodologies (such as segmentation), experiments and patient information (such as age, gender, handedness, education, and other) that is minimalist in nature. The user enters the information in his own secure system and the system translates this data through a client model to a MetaDL set of parameters which it then proceeds to evaluate internally. The integration of the metadata is supported by a toolkit that recognizes the type of object to be represented, thus offering the user a limited and structured set of options.

Basic searches provide a query term, such as a keyword or a sample image, that is looked up in the original data. In contrast, *advanced search* is a type of search patterned after expert boolean searches in Web search engines, but which is specific to a MetaDL. Since a MetaDL consists of meta-data items of various sorts, most with limited range of values (such as dates of entry or number of items), it is possible to customize a search ranking for the particular aspects a user is looking for. One might put "hard" limits on certain meta-data (for example requiring fairly recent entries) and put preferences on others (for example expressing a strong or weak preference for the number of items being large). The customized ranking r is formulated by assigning to each data item x a linear combination of meta-data "choice functions" c_i with associated *importance coefficients* α_i, where α_1, …, α_k are specified by the user and α_0 is a coefficient provided by the system to incorporate feedback from previous searches:

$$r(x) = \sum_{i=0\ldots k} \alpha_i c_i(x) \quad (1)$$

Choice functions are functions of meta-data that select a range on the particular ordering appropriate for each meta-data type; for example, the most common choice function for date is "newer than …", and the most common for size (number of subjects) is "of size at least …". The importance coefficients attached to each are discretized to a small number of values, each of which can be represented by a word or phrase for user input: "necessary" (1), "very important" (0.8), "important" (0.5), "not very important" (0.2), and "doesn't matter" (0).

For example, a user wants to find similar studies about '*volumetric analysis of hippocampus structure in schizophrenia using MRI scans and involving as many subjects as possible*' (Table 1). If the user is more interested in the "volumetric analysis", he/she can apply scheme 1 and system will return R1 as the best matched result. On the other hand, if he/she is more interested in the size of study set, scheme 2 can be applied and the system will return R3 as the best matched result.

III. A MetaDL EVALUATION METRIC

Evaluation of DLs [4] can be considered in the system (e.g., interoperability, scalability, heterogeneity, reliability, and integration), user (e.g., relevance, specificity, timeliness, effort vs. effect, and usability), and content (e.g., sufficiency, currency, and quality) domains. The evaluation addressed in this paper resides in the user and content domains; specifically, how closely MetaDL contents match the interests of users. We propose a measure of this match, based on the structure of advanced meta-information queries like those described for the BrassDL system in Section II.

Because a MetaDL by definition includes standardized meta-data for each item, searches involving these meta-data are easier to compare than in the general case. "Basic" searches, e.g. keyword searches, are not considered here. For advanced searches, prioritizations can be aggregated and considered independently of particular searches, resulting in an overall measure of which meta-data (and thus which data) are most relevant to most users. In BrassDL, for example, we might find that for "data type" fMRI scans are preferred to PET, or that age of data is (or isn't) an important factor for users.

The evaluation of data relevance as described is only an approximation, since it generalizes queries done with combinations of meta-data to each meta-data constituent independently of the others. A more refined variation of this is to consider sets of k meta-data items, $k \leq n$, where n is the total number in the system, in combination. For

Related Features	Analysis Method	Region of Interest	Disease	Num. of Subjects	Acquisition modality	
R1	Volume (c = 1)	Hippocampus (c = 1)	Epilepsy (c = 3)	20 (c = 3)	MRI (c = 1)	
R2	Shape (c = 2)	Hippocampus-amygdala (c = 2)	Schizophrenia (c = 1)	30 (c = 2)	MRI (c = 1)	
R3	Shape (c = 2)	Corpus Callosum (c = 3)	Schizophrenia (c = 1)	50 (c = 1)	MRI (c = 1)	
Importance Coefficient Set 1	1	0.5	0.5	0.2	0	∑ R1 = 3.6 ∑ R2 = 3.9 ∑ R3 = 4.2
Importance Coefficient Set 2	0.5	0.5	0.2	1	0	∑ R3 = 3.7 ∑ R2 = 4.2 ∑ R1 = 4.6

Table 1. Finding similar studies. For each related feature (first line), the three records listed (R1, R2, R3) contain a value and corresponding result from the choice function c (lower values indicate higher ranking). Two importance coefficient sets are given to show two different overall rankings (in the bottom right cells; again, lower values indicate higher overall ranking).

k = 1, this is identical to what was described above. For k = 2, it considers the relative importance users attach to *pairs* of meta-data items. Extending the example above, this might allow us to determine, for example, that PET data is very popular for studies of schizophrenia, and that old data is more often sought for cognitive tasks. The framework also provides a numerical ranking of each of these, which allows them to be compared to each other.

It is not reasonable to calculate every combination for every value of k, since the number of combinations increases factorially: $N_k = \frac{N!}{(N-k)!}$. However, it is still possible to consider and compare all combinations below a threshold value (e.g., k = 3), and it may be informative to determine the highest-valued combination for some higher values of k.

The formulas below calculate an R value that evaluates *how important* a metadata item (e.g. datatype) is on average. Let N be the number of advanced search queries, N_k the size of the search group that includes search k, G the number of search groups (an obvious use of this is one group per user, but can be any grouping), $\alpha_{i,j}$ the prioritization of metadata item i in search j.

Evaluation of a metadata item: $R(i) = \frac{1}{N}\sum_{s=1}^{N} \alpha_{i,s}$

Same evaluation with weighting for number of user queries, if available: $R^*(i) = \frac{1}{G}\sum_{s=1}^{N} \frac{\alpha_{i,s}}{N_s}$

Evaluation of a pair of metadata items: $R(i,j) = R(i)R(j)$

Evaluation of any number, k, of metadata items: $R(1,\ldots,k) = \prod_{i=1}^{k} R(i)$

Prioritizations of meta-data parameters in an advanced search can take on several different values, and so in the latter case is necessary to add a second parameter to the evaluation of data relevance that we calculate: a mapping for the importance coefficients allowed in expert search (Section III). With this second parameter, it is possible to formulate evaluations of popularity like the following: “what is the most common combination of meta-data requirements of 'very important' or higher” or “what is the most often required meta-data value”. It is also possible to add as additional parameters cutoff values for the expert search choice functions, resulting in the ability to evaluate “what is the distribution of user requirements for recency for different data”, or “what is the most preferred minimum size for PET datasets”.

The choice function cutoffs are necessarily different for every metadata item, since they all have different ranges (and data types: integer, floating point, or text). Because of this, we use a set θ of fixed cutoff parameters (one per metadata item) and a function f_θ that filters according to cutoff:

$$f_\theta(\alpha) = \begin{cases} \alpha & \text{if } \alpha > \theta \\ 0 & \text{otherwise} \end{cases}$$

Then the evaluation formula, *R*, is as before except for the application of the filter f_θ:

$$R(i) = \frac{1}{N} \sum_{s=1}^{N} f_\theta(\alpha_{i,s})$$

and likewise for *R**. This allows one to evaluate, e.g., "what are the most often required (α = 1) metadata items?" The result is a numeric ranking of metadata items, and as described previously, for each a histogram can be calculated giving the proportion of each value that was required (e.g., "PET" or "fMRI" for datatype).

IV. CONCLUSIONS

We have presented BrassDL, a MetaDL for human brain data that incorporates highly heterogeneous imaging data and results, and an evaluation metric that allows the user to easily narrow a search to those data sets that best match the user's needs.

MetaDLs are timely and important in the field of neuroscience to support research-activity documentation, scientific discovery, collaboration and sharing of extremely valuable data. As new imaging technologies develop, the field is exploding with new, highly valuable data that amass in individual laboratories with public funding and which are basically non-shareable. Access and combination of these data will lead to very large sets that connect the function and structure of human brain, furthering the goal of the Human Brain Mapping Project.

Evaluation of expert search popularity is a useful tool in evaluating how a MetaDL corresponds with its users needs.

REFERENCES

[1] F. Makedon, J. Ford, L. Shen, T. Steinberg, A. Saykin, Heather Wishart, and S. Kapidakis, "MetaDL: A Digital Library of Metadata for Sensitive or Complex Research Data", submitted to ECDL, Sept. 2002, Rome, Italy.

[2] Fox, P. T. and Lancaster, J. L. (2002). Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319-321.

[3] Grethe, J. S., Van Horn, J. D., Woodward, J. B., Inati, S., Kostelec, P. J., Aslam, J. A., Rockmore, D., Rus, D., and Gazzaniga, M. S. (2001). The fMRI data center: An introduction. *NeuroImage*, 13(6):S135.

[4] Ronald L. Larsen. The DLib Test Suite and Metrics Working Group: Harvesting the Experience from the Digital Library Initiative. *D-Lib Working Group on Digital Library Metrics Website*, April 2002. http://www.dlib.org/metrics/public/papers/The_Dlib_Test_Suite_and_Metrics.pdf

[5] METAe Metadata Engine. http://meta-e.uibk.ac.at/

[6] MUMIS: Multimedia Indexing and Searching Environment, University of Twente. http://parlevink.cs.utwente.nl/projects/mumis.html

[7] Dublin Core Metadata Initiative. http://www.dublincore.org/

[8] C. Houstis, S. Lalis, "ARION: An Advanced Lightweight Software System Architecture for accessing Scientific Collections", Cultivate Interactive, issue 4, May 2001

[9] J. Ford, H. Farid, F. Makedon, L. A. Flashman, and A. J. Saykin, "Classification of fMRI Brain Activation Maps", in preparation.

Evaluating User Behavior on Data Collections in a Digital Library

Michalis Sfakakis[1] and Sarantos Kapidakis[2]

[1]National Documentation Centre / National Hellenic Research Foundation
48 Vas. Constantinou, GR-11635 Athens, Greece
msfaka@ekt.gr
[2]Archive and Library Sciences Department / Ionian University
Plateia Eleftherias, Paleo Anaktoro, Corfu 49100, Greece
sarantos@ionio.gr

Abstract

We evaluate the usage of a Digital Library with many different collections, by examining its log files, and we concluded that the access points that the users mostly refer to, depend heavily on the type of content of the collection. We also found that most users not only tend to use simple query structures (e.g. one search term) and very few operations per session but they also reduce the complexity of their sessions, as they get more experienced.

Introduction

The evolution of Digital Libraries attends great interest by researchers in a variety of disciplines during the last years. Especially the study to understand and evaluate their usage has become a centric point in a number of Digital Library projects ([8], [9]) and specifies a number of critical factors during the design, creation and development process of a Digital Library ([10]).

Depending on the study and its use, a number of appropriate qualitative or quantitative methods exist ([3], [4]) to accomplish it. An unobtrusive way to study and evaluate user behavior is the Transaction Log Analysis. Although log analysis is used as an effective method to assess how users actually interact with a working Digital Library, this method hardly provides any information about the users' reasons behind their specific behavior - which is also very difficult to extract – and it is lack of giving information on their intentions. The accuracy of this quantitative method heavily depends on the detail of the information logged (automatically by the system), the period of time used to log the information, the usage and the number of the performed transactions during the log period. Such data are not usually publicly available (especially in detail) because of privacy constraints. For these reasons and due to that large Digital Libraries have recently started developing, only a few studies exist based on this technique ([5], [6], [7]).

In this work, based on the logged information, we study and evaluate the behavioral tendencies of different user groups on a variety of collections in the Digital Library of the Hellenic National Documentation Centre (*NDC*). The Digital Library of NDC (http://theses.ndc.gr) is one of the most significant in Greece and consists of more than ten collections of diverse types. Most of these collections are unique world wide with internationally interesting content. In particular, the "Hellenic Ph.D. Dissertations Thesis" collection is part of the international Networked Digital Library of Theses and Dissertation Initiative ([2]). The Digital Library of NDC is targeted to a number of diverse types of user groups (e.g. students, researchers, professionals, librarians, etc.), mainly in Greece, from a variety of scientific domains.

In the following section we describe the goal and the methodology of this study. We also describe the collections, their characteristics, the target user groups they refer to and the functionality of the available operations by the system. Then we present some our most important observations from the search operation usage and formulation, the Access Points usage and how users accomplish their requests, together with our interpretation and conclusions. Finally we present a number of interesting issues arrived from this work for further evaluation and research.

Purpose and Methodology of the study

The goal of this study is to compare and evaluate the differences on the usage among data collections, based on the collection content type, metadata and characteristics and also to approach the way diverse kinds of users accomplish their requests.

For a period of twenty months, we logged the operations performed by the users on the content of many different collections of a Digital Library, using a specific web based retrieval system. Considering the content type (e.g. PhD theses, articles in a specific scientific area, Books and Periodicals Union catalogues etc.), the structure and the quality of the collection, plus the target group they refer to, we selected the ten most used ones and classified them into four categories.

Category one consists of the collections: *Hellenic Ph.D. Dissertations Thesis* (*C1*) and *Hellenic Scientific Libraries Serials Union Catalogue* (*C2*), targeted to diverse kinds of scientific user groups (e.g. students, researchers etc.) from all scientific domains.

Category two consists of the collections: *Medical Bibliography – Hippocrates* (*C3*) and *Social Science Bibliography – GLAFKA* (*C4*), with simple metadata structure, targeted to a specific scientific user group (e.g. doctors, sociologists, researchers).

Category three consists of the collections: *Hellenic Archaeological Records – grARGOS* (*C5*) and *International Archaeological Records – intARGOS* (*C6*), including library material with diverse types of data, targeted to a specific scientific user group (e.g. researchers on Archaeology).

Category four consists of the collections: *Hellenic School Libraries* (*C7*) and *Hellenic Public Libraries Union Catalogue* (*C8*), union catalogs for library materials from many domains, targeted to librarians.

The remaining two collections are the *ARGOS – Serials Union Catalogue* and the *Evonimos Ecological Library*, which enjoy smaller use and we do not examine them separately, for simplicity, but we count their usage on the aggregated results.

All the above collections are structured using the UNIMARC format but they do not use the same detail on metadata description. From their 300,000 metadata records there are links giving online access to 14,000 digitized documents composed of 2,000,000 scanned pages and few other object formats.

The web-based retrieval system that we monitored is implementing a Z39.50 client and connects to a Z39.50 server. The users start their sessions by selecting and connecting to a collection. After connecting to a collection, a user may express his search request or browse specific *Access Points* (e.g. extracts information about metadata indexing - the terminology used for naming them is the one of the Z39.50 attribute set bib-1 as defined in [1]) and then to retrieve (*present*) the documents. In some cases, there is the ability to further access the object (document) that includes the full text, mostly in scanned images. There are seventeen available Access Points and the "search" operation supports Boolean combinations of them. When the user browse the terms from a specific Access Point, the system permits either to select a term in order to use it in a "search" operation or to retrieve ("present") the corresponding documents for display or further processing (searching and retrieving). From the more advanced searching techniques, the system also supports Boolean search combination of previously issued result sets, Search History and Selection of specific records from individual result sets.

The study covering period from August 2000 till March 2002, gave a set of 490,042 operations to process and evaluate (Table 1).

Table 1. Summary of Processed Data

Covering Period	Number of Collections	Number of Sessions	Number of Operations
August 2000-March 2002 (20 Months)	10	64,597	490,042

Fig. 1 and Fig. 2 depict the number of operations and sessions respectively for all collections on a monthly basis. During the study covering period, no major modifications occurred on the two basic components of the Digital Library, the collections and the retrieval system. On the other hand, the number of operations (Fig. 1), sessions

(Fig. 2) and users increased while maintaining a yearly periodic variation, and the number of sessions and operations seem to have the same transitions.

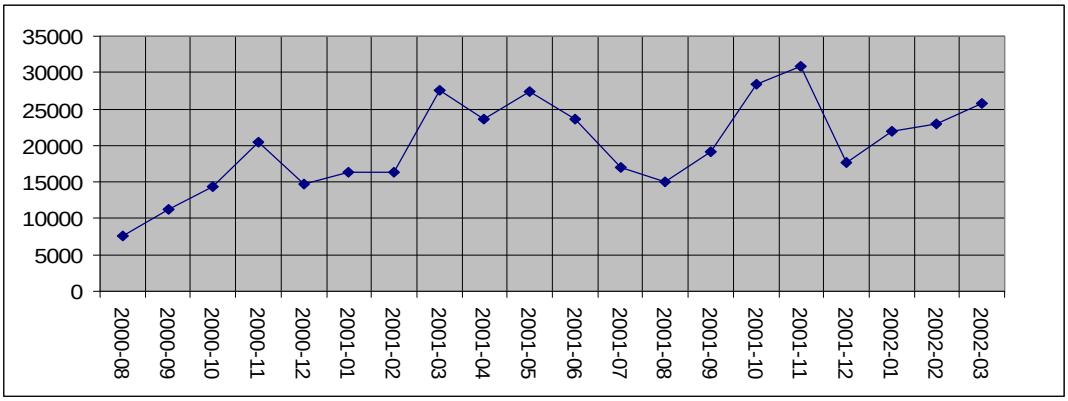


Fig. 1. Number of Operations per month

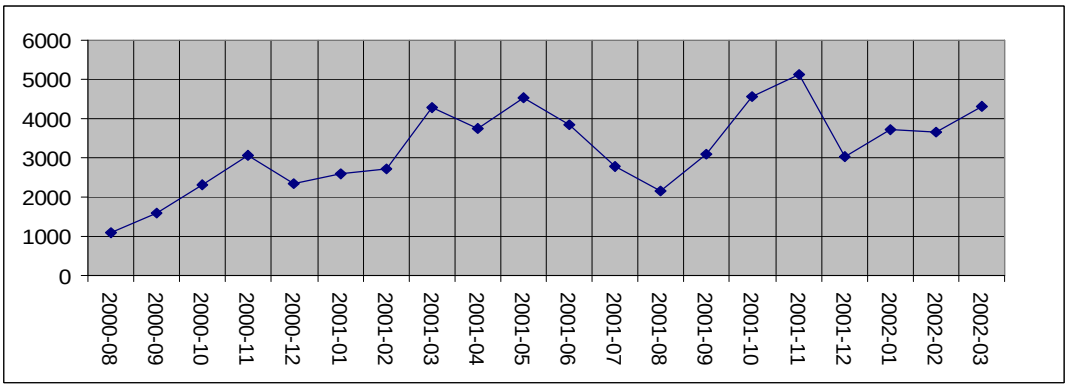


Fig. 2. Number of Sessions per month

Search Operation Usage and Formulation

From the processed logs, as presented on Table 2, we found that the percentage of the search operations is 38.34% of the total number of operations. The majority (81.75%) of these search operations were formulated using one search term (simple query). Finally, the users did not make use of advanced querying techniques (0.1%) by formulating search operations using previously issued result sets.

Table 2. Summary of Search use and formulation

Total Search operations	187,898 (38.34% total operations)
Search operations with only one search term (Simple Queries)	153,283 (81.57% Search Operations)
Search operations using Boolean expressions (Compound Queries)	34,615 (18.43% Search Operations)
Use of previously Issued result sets in compound queries	200 (0.1% Search Operations)

Use of Access Points Evaluation

Table 3 shows the number of times each Access Point has been used, for each collection and all collections together, and also the Access Points order of preference. The number after the Access Point name is from the Z39.50 bib-1 attribute set.

The evaluation of the usage of Access Points, shows that the most commonly used Access Points, for all the collections in the Digital Library, are the "Any", "Author", "Title", "Subject Heading" (Table 3), from the seventeen ones used in the metadata (the used terminology for naming Access Points is the one used by the Z39.50 attribute set bib-1). The vast majority of all users, independent of user group, used the "Any" Access Point for almost all collections. The only exception occurs at collection C6, were the most used Access Point is the "Author" which could be explained from the specialized subject area of the collection's content (Archaeological Records) in combination with the specific type of its closed targeted user group's requests.

Another interesting observation with regard to the first two categories of the collections is the big usage difference (60.5% - 80.9%) between the two most used Access Points. These collections consist of content with simple metadata structure and are targeted to a number of diverse types of occasional users.
At the third category which consists of collections with typical library material (e.g. more complex metadata structure, diverse kinds of material) that impose a more accurate process by professionals with consequence a better quality of metadata and targeted to a more specific user group, there is a balance between the three most used Access Points.

The most balanced usage between the commonly used Access Points, happens at the forth category, which consists of collections with common characteristics as those in the third category and targeted to librarians.

Consequently, we observe that the usage of these commonly used Access Points depends mainly on the collection they belong to as well as on the user group type they are targeted to.

Table 3. Summary of Access Points use per Collection

	Total	C1	C2	C3	C4	C5	C6	C7	C8
Any(1016)	114,412	55,137	4,946	20,474	9,119	5,688	3,729	5,488	5,539
Author(1003)	43,031	21,823	770	1,446	1,061	4,093	4,445	4,236	4,146
Title(4)	37,048	14,227	1,942	2,679	2,338	3,802	3,252	3,861	2,810
Subject Heading(21)	22,760	10,997	594	1,073	657	572	641	3,883	2,454
		Any Author Title Subject	Any **Title Author** Subject	Any **Title Author** Subject	Any **Title Author** Subject	Any Author Title Subject	**Author Any** Title Subject	Any Author **Subject Title**	Any Author Title Subject

Table 4 displays the usage of Access Point combinations for each collection and all collections together. We first observe that the Access Point "Any" is not that dominant in Access Point combinations as it was in single Access Point specifications. We also observe that the difference between the two most used Access Point combinations follows the previously observed Access Point usage pattern. Finally, for the majority of the collections, the most commonly used combination of Access Points is the "Title-Any", except for the collections C5, C6, C8. We have already seen (Table 3) that these collections have a more uniform usage on their single Access Points, without overusing the "Any" Access Point, and consequently the most commonly used combination of Access Points for them is the "Title-Author".

We also observe that the Access Point "Title" is used much more often on Access Point combinations, although it is the third one in the list of the most used Access Points, which indicate that "Title" is used in more sophisticated "search" operations and by more sophisticated users.

Table 4. Summary of Access Points Combination use per Collection

	Total	C1	C2	C3	C4	C5	C6	C7	C8
Title(4) - Any(1016)	4,262	**1,925**	**137**	**640**	**489**	177	236	**370**	179
Title(4) - Author(1003)	2,860	776	64	107	43	**656**	**541**	294	**301**
Author(1003) - Any(1016)	1,388	823	51	104	41	69	119	54	109
Subject Heading(21) – Any(1016)	976	540	40	86	47	33	37	79	77
Title(4) – Subject Heading(21)	602	286	24	33	21	25	35	67	70
Title(4) - Author(1003) – Any(1016)	506	293	18	39	24	22	30	15	50
Subject Heading(21) - Author(1003)	503	264	10	21	10	23	28	60	66

Comparing the results that, the vast majority (81.57%) of the search queries consist of one search term (Table 2) and most users for almost all collections use a general Access Point ("Any") to accomplish their requests with big usage difference from the next, more specific, Access Point ("Author or "Title""), we can derive that new users will need more operations to accomplish their requests which impacts the increase of the number of operations per session when new users enter the system.

User Behavior (how user accomplishes the job)

Fig. 3 shows the monthly average operations per session aggregated for all ten collections studied, on a monthly basis. Similar lines correspond to each one of the studied collections.

The average number of operations per session in general drops during the study period. Does this mean that the vast majority of old users becomes more experts and expresses their requests using fewer operations?

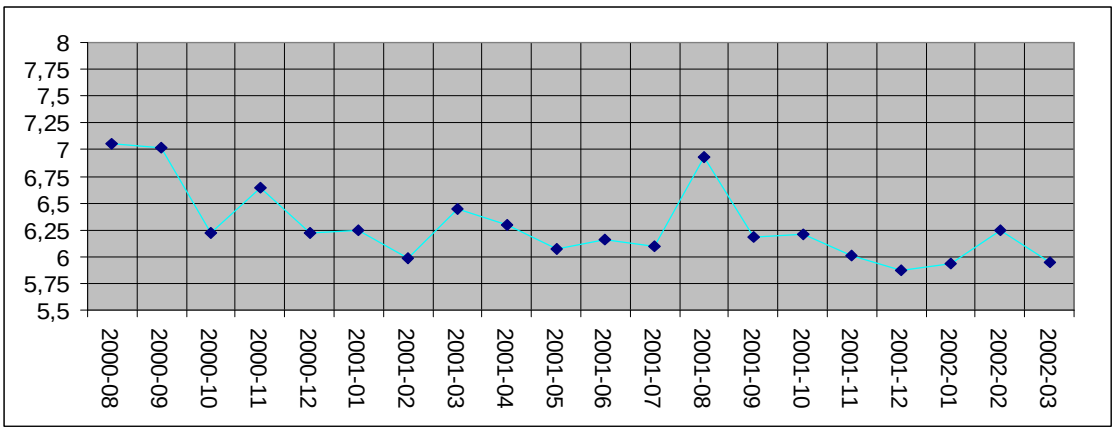


Fig. 3. Average Operations per session

Fig. 4 shows the monthly proportion of sessions with operations less than or equal to three per session aggregated for all ten collections studied, on a monthly basis. Similar lines correspond to each one of the studied collections.

We observe that in each month (Fig. 4), three operations are enough to fulfill at least half of the sessions. Also, the number of sessions with less than or equal three operations per session have a constant fluctuation, which indicates that there is a balance on the number of sessions with the same number of operations per session, between new users and old users that become more experienced.

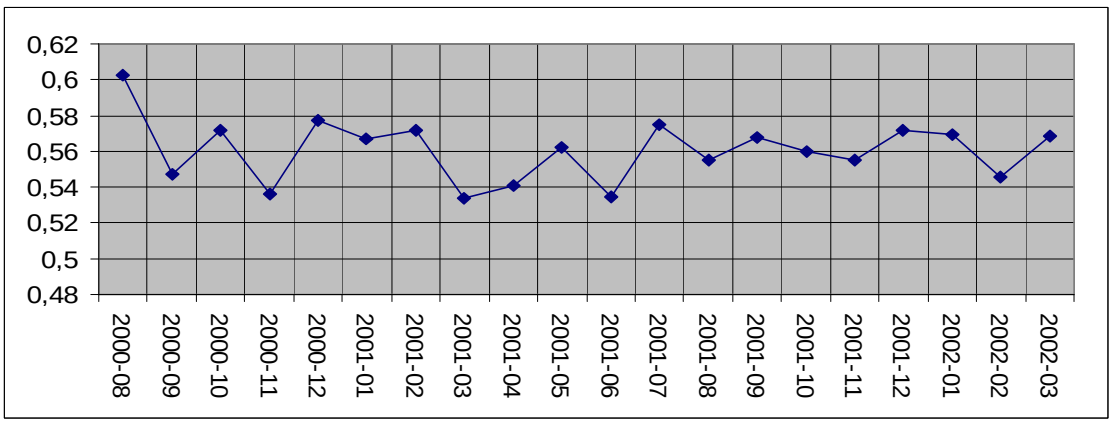


Fig. 4. Proportion of Sessions with operations less than or equal 3 per session

Another interesting question is how we measure the experience of the users. The experience of the users will certainly increase by time, but how can we distinguish it from that of newer users, on a system that does not record the identification of the users?
We assume that one aspect of the experience of the user is measured by the number of operations that are included in a session, the full sequence of operations that the user performed. We have already concluded that most users perform few operations in order to find their material, but as the users become more experienced, do they use more operations (been able to make more complex sessions) or less operations (been more specific and efficient) in their sessions? The addition of new users into the system makes the distinction more difficult.

Fig. 5 shows the number of sessions for each number of operations (from 1 to 30) per session for five representative months, aggregated for all collections. From fig. 5 we can see that on the later stages in our Digital Library lifetime, the increased number of users corresponds to only an increase to the number of sessions that have only one operation. We already observed, on the evaluation of Access Points usage, that new users perform queries with many operations per session. We also believe that it is unlikely that all new users perform only queries with one operation per session, while we can see from fig. 2 that the total number of sessions in the last three of the depicted months are practically the same, so we conclude that older users decrease the number of operations into their sessions, in a way that (by coincidence) corresponds to or outperforms the increase of new users performing the same number of operations per session.

Thus expert users use fewer operations per session than non-expert users, and the users decrease the number of operations in their sessions during the time they use the Digital Library.

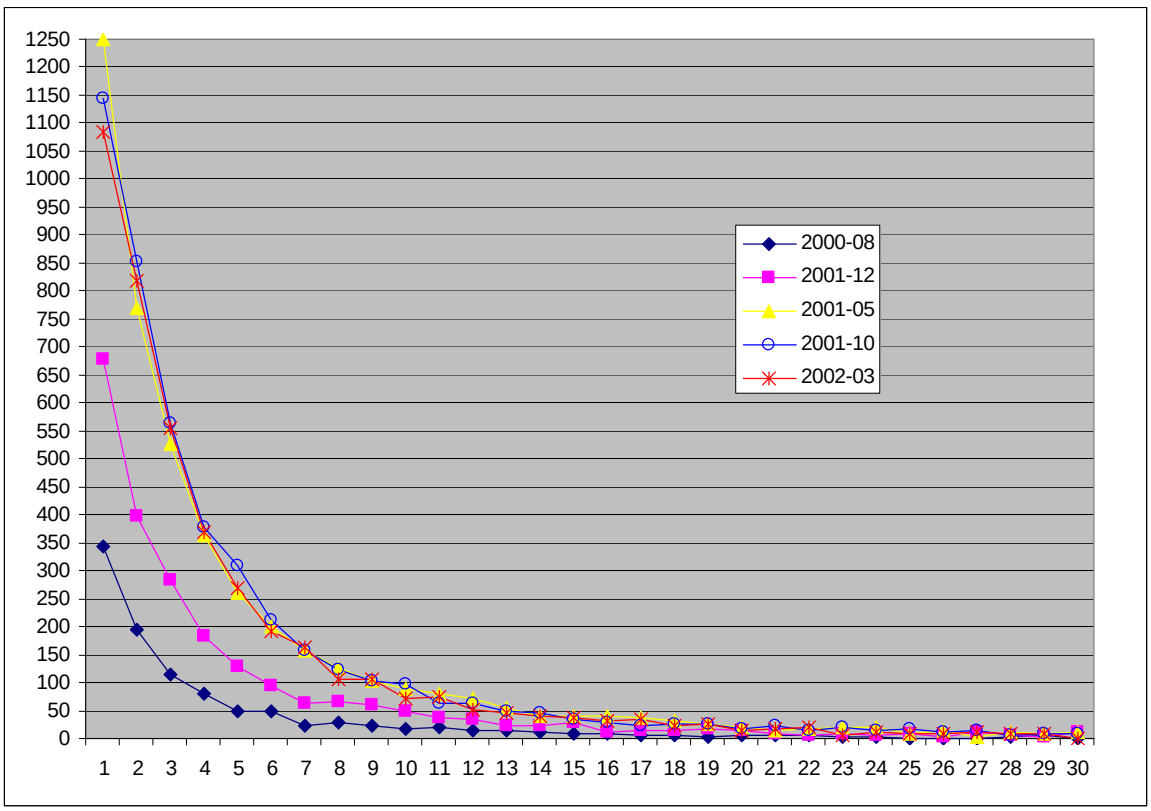


Fig. 5. Number of operations per session - Number of Sessions

Conclusions and Future Research

We studied the Access Points usage and we derived that the “Any” Access Point is used by novice or non specialized users, while other Access Points, like “Title” and “Author”, are mostly in use by experienced and sophisticated users, on complex queries and on collections with more complex metadata. We also examined the number of operations and sessions and we concluded that expert users tend to decrease the number of operations in their sessions, by been more explicit and efficient.

From this work a number of interesting points arrives for future evaluation and research. How does the Access Points usage evolve by the time? A more detailed analysis for the search term formulation (e.g. word, phrase, truncation) used by the same group of users to accomplish their search requests per collection would be interesting. How previously issued user behavior results, differentiated per collection? Another point of interest is how different user types (e.g. professionals, ordinary users) behave under the same circumstances. What sequences (patterns) of operations (i.e. number of "Presents" follows the "Search" operation, etc.) in sessions do different types of users adopt? Finally the Query formulation complexity progress during the time period is also another interesting point of evaluation.

References

1. ANSI/NISO: Z39.50 Information Retrieval: application service definition and protocol specification: approved May 10, 1995.
2. E. Fox, Robert Hall, Neill A. Kipp, John L. Eaton, Gail McMillan, and Paul Mather. NDLTD: Encouraging International Collaboration in the Academy. In Special Issue on Digital Libraries, DESIDOC Bulletin of Information Technology (DBIT), 17(6): 45-56, Nov. 1997.
3. Bains S., "End-User Searching Behavior: Considering Methodologies", The Katharine Sharp Review, No. 4, Winter 1997.
4. Covey, D. T., "Usage and Usability Assessment: Library Practices and Concerns", Washington, D.C., Digital Library Federation Council on Library and Information Resources, January 2002, ISBN 1-887334-89-0.
5. Jones, S., Cunningham, S. J., McNab, R. J. and Boddie, S., "A Transaction log Analysis of a digital library", International Journal on Digital Libraries, v. 3:no. 2 (2000), pp. 152-169.
6. Mahoui, M., Cunningham, S. J., "Search Behavior in a Research-Oriented Digital Library", ECDL 2001, LNCS 2163, pp. 13-24.
7. Mahoui, M., Cunningham, S. J., "A Comparative Log Analysis of Two Computing Collections", Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL Lisbon, Portugal, Sept. 2000, pp. 418-423.
8. Peterson Bishop, A., "Working toward an understanding of digital library use: a report on the user research efforts of the NSF/ARPA/NASA DLI projects", D-Lib Magazine, October 1995.
9. Payette, S.D. and Rieger, O.Y. "Z39.50 The User's Perspective", D-Lib Magazine, April 1997.
10. Van House, N.A. et. al., "User centered iterative design for digital libraries: the Cypress experience", D-Lib Magazine, February 1996.

Virtual Agents for a Bookstore: an Empirical Evaluation

Pasquale Lops[1], Verner Andersen[2], Hans H.K. Andersen[2], Fabio Abbattista[1], Giovanni Semeraro[1]
[1] Dipartimento di Informatica, Universitá di Bari, Italy
e-mail: lops@di.uniba.it; fabio@di.uniba.it; semeraro@di.uniba.it
[2]Risoe National Laboratory, Denmark
e-mail: verner.andersen@risoe.dk; hans.andersen@risoe.dk

Introduction

According to Lesk, "a digital library is not merely a collection of electronic information" [10]. It is "a distributed technology environment that dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration and reuse of information by individuals and groups" [11]. As a consequence, digital libraries can play a relevant role in several key areas of the *e-era*, such as *e*-government, *e*-learning, *e*-publishing, and *e*-commerce. Specifically, in the field of e-commerce it is well known that the process of buying products and services often implies a high degree of complexity and uncertainty about the conditions of information seeking, about items for sale, the purchase of wanted products and the actual navigation on a site. Some important problems concerning e-commerce in general and shopping at Internet bookstores in particular are (for more information see [3,4]): Getting people started on the Web and making their first purchase; using traditional metaphors for shopping on web sites; users are so to speak forced to make their model of shopping fit into a web structure with which they are not familiar; getting people to submit personal information.

The common theme for the mentioned problems is *uncertainty*. Uncertainty about the new media, the new ways of shopping, adequate representation of products, trust in the e-commerce sites, the navigation of specific sites and the actual procedures for buying. It is very important to overcome these problems in order to facilitate the use and acceptance of e-commerce. The solution we proposed for the COGITO[1] project is based on "intelligent personalized agents" (chatterbots) which represent virtual assistants or advisors (also visually) by modelling their ability to support customers. There are many possible applications for such virtual assistants. They could instruct customers in the use of a Web site, point out new offers, help sift through products, and other support. The main problem of most of today's Web services is that they offer manifold navigation options and (usually simple) search functions, but leave it up to users to find their way through the many interface functions, understand them and interrelate them cognitively. Usually, users have to decide themselves which sequence of actions must be performed to solve a given task. Complex search queries, for example, must be constructed step by step. Beginners and occasional users are often daunted by the complexity of today's services and thus need "pro-active" support or advice from the system in order to fully utilize the range of functions available. In order to verify the assumptions underlying the design decisions above, and to find out appropriate ways to adjust the system parameters, the technical development is accompanied and heavily influenced by in-depth evaluations of both the individual components as well as the system as a whole.

Profile extraction to tailor contents: advantages

Personalization is very common in the area of e-commerce, where a user explicitly wants the site to store information on herself, such as her preferences. In fact, the more a system knows about users the better it can serve them effectively. But there are different styles, and even philosophies, to teach computers about user habits, interests and preferences.

User modelling simply means ascertaining a few bits of information about each user, processing that information quickly and providing the results to applications, all without intruding upon the user's consciousness. The final result is the construction of a user model or a user profile [8]. By user profile we mean all the information collected about a user that logs to a web site, in order to take into account her needs, wishes, and interests. A user profile, as intended within the COGITO project, is composed by two main frames - the frame of user data, which comprehends interaction data (number of searches or purchases within a category, number of connections, etc.) and the frame of the user interests, which is a part of the profile built on the basis of supervised learning algorithms [1]. The Profile Extractor module is the personalization component which dynamically discovers user preferences from data collected in the course of past dialogues with the chatterbot. By examining the dialogue histories, it extracts some characteristics that are useful for recognizing the categories preferred by a buyer. The preferences for the user automatically "learned" by the system concern the ten main book categories the BOL (Bertelsmann Online virtual shop) product database is

[1] *COGITO - e-Commerce with Guiding Agents based on Personalised Interaction Tools - is an EU-funded project in the 5th Framework Programme, (IST-1999-13347)*

subdivided into. User profiles are represented by XML files and are the key to personal recommendations because they enable the agent to customize its recommendations to the individual user.
As pointed out in [14], the main advantages of using this approach in e-commerce are:

- making the site more attractive for users: A web site that takes into account user preferences is able to suggest products reflecting customer needs. It will probably turn a significant part of browsers into buyers;
- obtaining customer trust and confidence: Users will not be requested to explicitly insert information concerning their preferences and tastes, but they will be able to participate in the management and updating of their personal profile. This will result in an increase of their trust and confidence in a system able to automatically collect data about their preferences;
- improving customer loyalty: the effectiveness of a personalization system improves in the long run. Every time a customer interacts with the web site, the personalization mechanism collects new data about her preferences, so that a more and more satisfactory service can be offered. In this case, passing to the competition is often unfavourable for a customer. In fact, even if a competitor uses a personalization system, it has to learn a lot of information about the new customer to be able to offer the same satisfactory service.

The retrieval process in COGITO

On the ground of user profiles, the chatterbot offers a better support to customers during the interaction, providing personal recommendations, purchase incentives and helping users in problematic situations during the search. This improves the usability of the BOL web site, as shown by the following scenario.
A user is known by the COGITO system if she completes the BOL registration procedure. This step allows to provide each customer with a personal identification number. It is necessary to recognize a user accessing to the on-line media shop and collect data about her preferences for generating/updating her profile.
A dialogue between the chatterbot, named Susanna, and an unknown user asking the chatterbot for a book by author "King" is undertaken:

Susanna:	Good evening! Nice to see you found your way to BOL.DE. May I introduce myself? My name is Susanna and I am your personal assistant at Bertelsmann Online. What is your name?
User:	My name is Fabienne.
Susanna:	Nice to meet you, Fabienne. What can I do for you?
User:	I'm looking for a book.
Susanna:	Then this is the right place to be! We here at BOL have hundreds of thousands of deliverables books. Are you looking for a book by a specific author?
User:	Yes
Susanna:	What is the name of the author?
User:	King
Susanna:	I shall check whether we have a book by the author called King. One moment please…

Susanna finds several books by the author "King" through a remote call (deep linking) to the search engine available on the BOL web site and displays them, as shown in Figure 1a.
It can be noticed that the books ranked first are by the author Stephen King. Books by other authors are found further down the list, which means that the user should scroll down a long list if she was not looking for a book by Stephen King. The customer not looking for a Stephen King book might now also choose to either refine the search by using an advanced search function or continue to chat with Susanna. If the user has already been chatting with Susanna about some of her interests, then a profile of this user is available to the system, which can exploit it to accomplish a more precise search in the product database.
Let us suppose that the profile of such a user contains the category "Wissenschaft_und_Technik" (Science and technique) as one of the preferred categories with degree of preference 100% (see Figure 2), therefore the result of the query shows books in this category at the top (Figure 1b). In this case, the system exploits a query expansion mechanism to modify the original query "King" into "King" AND "Naturwissenschaften", a subcategory of Wissenschaft_und_Technik. In fact, when a user asks the chatterbot for a book by author "King", it dynamically builds an XML file containing the string "King" as the value of the proper tag – in this case <author> – and sends it to a module called Prompter [9], that performs the expansion of the original query by using the favourite book categories stored in the profile of the user. The query expansion process consists of an improvement of the criteria used for the specification of a query. This is usually achieved by modifying an already defined query by adding search terms taken from three different sources: 1) Product Thesaurus - products like books or other kinds of media are usually characterized by a textual description. The most relevant words contained in these descriptions are clustered according to their relation to the most frequently appearing ones, thus generating a "thesaurus" of terms; 2) User Profiles - they are accessed for identifying the book categories preferred by a user, which can be enclosed in a query for a more specific result identification, as already described in the previous section. 3) Usage Patterns - the application of association rules to a specific user can lead to infer a possible interest of this user in a product or service [2].

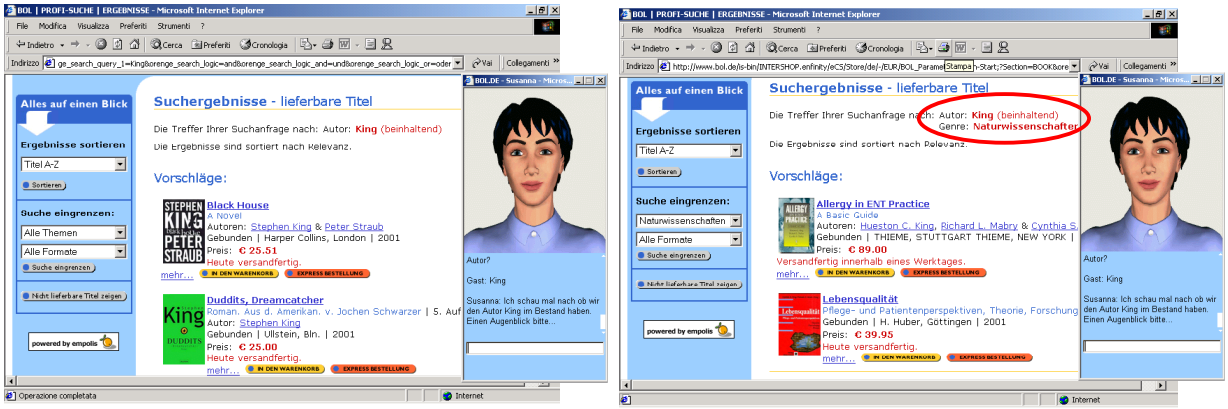


Fig. 1. (a) Susanna offers a long list of books belonging to several categories by authors whose last name is "King". (b) List of books by authors whose last name is "King" and belonging to the book category "Naturwissenschaften" (in the red circle).

In this way, the chatterbot can decide which dialogue context to use when the dialogue comes to a dead end, i.e. when the user does not want to neither take the initiative nor mention a specific topic of discourse. This scenario highlights the dependence of the result set on the profile of the user that issued the query. The information retrieved by the interface to the information sources (i.e. the expanded keywords, the preferred book category or the dialogue context coming out from an applied usage pattern) is used for the generation of a deep linking, which is directly forwarded to the chatterbot by the Prompter. The Prompter is responsible for determining a suitable query expansion method to be used, according to the information available in the input file. The decision process is thoroughly described in [9].

Empirical evaluation

The framework used for evaluating the performance of the agent is based on the means-end hierarchy already utilised during the phase of requirement specification (see Andersen et al. [3]). For further information about the means-end hierarchy see Rasmussen et al. [12]). In this framework the requirements are classified in three levels: the strategic-, the procedural-, and the operational level. This structuring will furthermore facilitate the three levels of evaluation of the COGITO outcome as follows:

- verification, which is a check of implementation of operations specified in the user requirements, and therefore directly related to the lower level of the hierarchical representation of user requirement, the operational requirements;
- evaluation, which is a check of the presence of the functionality specified in the user requirements, i.e. is the system capable of executing all the sequences of operations needed for fulfilling the goals specified in the requirements. This part of the test procedure is directly related to the middle part of the hierarchical presentation of the requirements, the procedural requirements;
- last, but not least, the validation takes care of - based on user satisfaction - testing whether the system is of any value to the end users, i.e. do they perform better, more efficient and with a higher success rate than without having the system available. The question here is the difference between developing the system right, i.e. following carefully all the elicited requirement specifications, or developing the right system, i.e. a system that really is of benefit to users.

The logical way of evaluating a system is to take the top-down approach in which the evaluation and validation process is tested by user interaction with the system. In case this test does not end up satisfactorily,

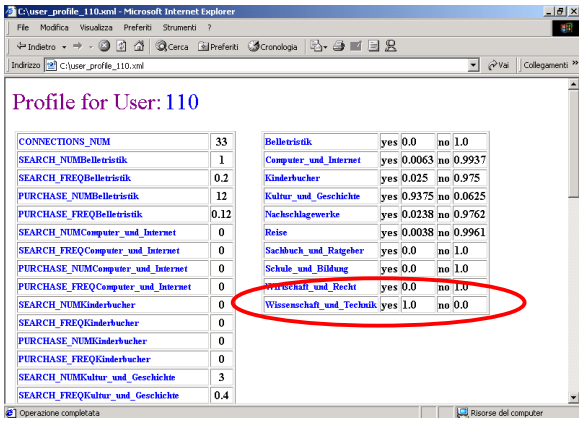


Fig. 2. The profile of user #110.

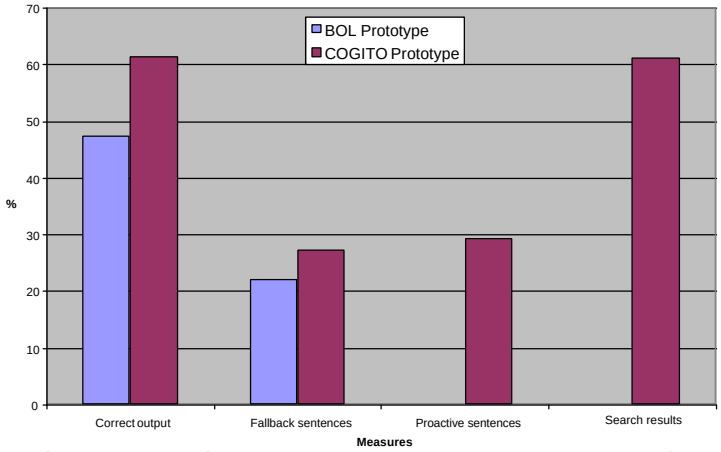


Fig. 3. The result of the analysis of the agent-user conversation log in terms of number of correct text output, fallback sentences, and pro-active sentences. In addition, the figure shows the agent performance with respect to the percentage of positive search results.

the next step would be to take the bottom-up approach starting with the verification phase checking the implementation of operational features and continuing with check of the functional features. That is, the COGITO system evaluation is based on the top-down approach focusing on the system/user interaction. In more detail the 'evaluation and validation' of the COGITO agent has been performed by letting groups of test persons solve various tasks related to searching general information or specific products utilising the agent and its linking with specific BOL sites related to the questions and wishes of the users. The evaluation is partly based on quantitative measures related to analysis of the conversation log, partly on eye-tracking specifying the time the user spent looking at the agent, the answers given by the agent, or the BOL site itself, and partly on qualitative measures based on fulfilment of detailed questionnaires. The test persons were requested to complete the questionnaire revealing their satisfaction with the system and the agent concerning various aspects, such as impression, control, effectiveness, navigability, learnability, aidability, and comprehension of the agent.

In order to have a reference for evaluating the COGITO proactive agent, a baseline session was performed using the BOL site equipped with a 'BOL state-of-the-art agent' for comparing the two agents (see Figure 4). This agent had a level of chatting performance in line with existing agents of today and was integrated to the BOL site by having simple links to products being requested by the customer, i.e. this agent had no proactive features.

Conversation Log Analysis

The analysis of the conversation log served the purpose to measure the conversation performance in terms of number of correct text output, fallback sentences, and pro-active sentences (see Figure 3). The measure "Correct text output" is based on manual analysis and interpretation of successful elements of the agent-user dialogue consisting of one user text input string, e.g. a request for information or a search query, and one agent output text string, e.g. delivering a correct answer and/or requesting further information from the user. The COGITO agent shows a better performance than the BOL agent with respect to the correct output

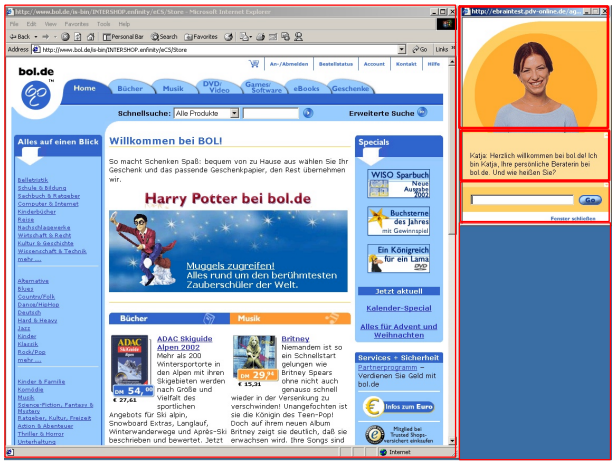


Fig. 4. A snapshot of the BOL site and the agent, as they appeared during the evaluation of the BOL agent. The red color "boxes" (dark for black and white copies) mark the examined areas of interest.

category (61% vs. 47%). It seems that the COGITO agent is better at recognizing search terms; it has a better vocabulary and rule set than the BOL agent. Furthermore, the fact that the BOL agent is relatively passive and therefore does not in the same sense as the COGITO agent produce true pro-active requests in response to user input adds further to the explanation. A pro-active agent sentence is counted every time the agent successfully takes a word or a string of words and constructs a contextually meaningful response to user input.

The BOL agent of course also asks questions, but they are more passive and general, and the agent does not utilise user input in stating further requests. These types of sentences we do not count as proactive.

In addition, we have also analysed the heterogeneousness of the conversations using two measures: the proportions of fallback sentences and the various sentence categories applied by the agents. We consider a large occurrence of fallback sentences as an expression of poor conversation performance. That is, there is a risk that the users will get frustrated if the agent "speaks" in set phrases or clichés, which means that the users experience a rather stereotypical conversation. In this sense, both agents did not show optimal performance. The BOL agent performed a bit better than the COGITO agent (22% vs. 27%), but one has no feeling of a significant difference between the two. In addition, the COGITO agent showed a larger variation in the conversation by using 9 different categories of fallback sentences while the BOL agent only used 6 different categories. Moreover, we have also looked at the search results generated by the agent in prompting the BOL search machine. In this respect a successful query is counted every time the agent on the basis of selected user input plus the added search terms created by the query expansion process prompts the BOL search machine with queries that produce a correct list of search results in terms of relevance for a given user situation (task) no matter if the user recognizes this. That is, we have repeated all the search queries listed in the conversation log using the BOL search machine and analysed the result in relation to the users tasks. The non pro-active BOL agent does not produce search queries. Instead, it operates with a concept of static "deep linking" based on general input from the user. However, most static deep links did not function because BOL changed its platform during the evaluation session. Furthermore, we have analysed the average length of user queries. On an average, a user query contained 5,05 words. In an analysis of queries posed by users at Excite, a major Internet search service, Jansen et al. [7] found that web queries are short. On an average, a query contained 2,21 words. Our users created queries that contained more than twice the number of words compared with the Excite users.

Nevertheless compared to studies of queries in traditional information retrieval systems like online databases (e.g. DIALOG) and public access catalogues (e.g. library catalogues) the COGITO queries are relatively shorter. In traditional information retrieval systems the queries on an average varies from 7 to 15 words dependent on the users expertise varying from novices to very experienced (for more information on this topic see e.g. Fenichel [6] and Spink and Saracevic [13]). In fact, one of the main ideas of introducing intelligent agents on the web is exactly to overcome some of the obstacles of traditional information retrieval like e.g. the use of Boolean operators and in the same way to allow users in natural way to type their queries in a conversational manner. With an average query length of 5,05 words, the pro-active agent tended to perform better than the traditional web based search engines (like the ones on Excite) without demanding the users to use any Boolean operators, and as shown in Figure 3 with relatively good performance in terms of search results.

Eye-tracking Analysis

We have used a remote eye-tracking system to measure the respondent's visual behaviour during the evaluation session. This device is non-intrusive and the respondents can behave as they normally would in front of a computer display. The eye-tracking system samples the eye-movements at 50 Hz to a data file. In addition, we have video recorded the eye-movements together with the graphic signal from the computer. The data is sampled only during respondents task solving. That is, we have recorded data from the point in time where the respondent is finished reading the task situation out loud and until she has finished the task, gives up or is stopped by the moderator. We have divided the screen into 5 so-called "Areas Of Interest" (AOI, see Figure 4) and named them 1) the agent torso, which shows the animation of the agent; 2) the agent text output field, where text from the agent is displayed; 3) the user input field, where the user can type, e.g., requests for information; 4) the BOL site where, e.g., search results are displayed; 5) the right lower corner, that shows the background of the screen. We have then calculated the amount of visual attention paid to each of these AOI in percentage of all viewing time during the task situations (see Figure 5). The smallest amount of viewing time has been spent looking at nothing at the right corner of the display. This is rather common to normal viewing behaviour when people is either daydreaming or solves cognitive task that do not require any new visual input. In addition, a small amount of viewing time has been spent on the visualisation of the agents. More visual attention has been paid to the BOL agent visualisation compared to animation of the COGITO agent (2,9% vs. 1,6%). The BOL agent animation has attracted approximately double as much visual attention than the COGITO agent. This is probably due to the more the photo-like appearance, the

obliging attitude, and a larger repertoire of gesticulations compared to the cartoon-like appearance of the COGITO agent. With respect to the agent part of the system, most viewing time has been spent at the text output field. This is not surprising that it takes time to read the text. In addition, some users had to scroll back to read longer paragraphs since the text "ran" to fast. There is only a 1,8% (14,4% vs. 12,6%) difference between the two prototypes with respect the text output AOI. A larger difference is measured in terms of the user input field (9,1% vs. 4,1%). It seems like the BOL prototype respondents used more time in checking their keyboard strokes than the COGITO respondents. This result is even more distinctive because of the average number of words typed by the BOL respondents compared to the COGITO respondents. The BOL agent's larger amount of mode errors and output errors may also play a role. That is, the BOL respondents wanted to be sure not to make any "typos" making the moderator think that the agent errors were their fault. Another explanation is, that the two groups differ with respect to their typing abilities. The analysis of the profiles of the two groups seems not to indicate any difference on this matter. With respect to visual attention at the BOL.DE site there is only a little difference between the two prototypes (21,3% vs. 18,6%). One explanation could be that most BOL agent deep links did not function well enough, which more often lead the BOL respondents to use the BOL site on its own to deal with a given situation than was the case with the COGITO respondents. In addition, the fact that the BOL agent deep links sometimes lead to rather surprising pages could play a role. If a respondent e.g. asked on how to use a credit card she was linked to a page that promoted a wide range of colourful cheap Christmas offers.

In general, all groups of respondents used approximately half of their display viewing time on the agent and half on the BOL.DE site. The BOL group of respondents looked at the site 21,3% of their display viewing time and 26,4% at the agent. The COGITO group looked at the site 18,6% of their display viewing time and 18,4% at the agent.

For both prototypes, much viewing time has been spent outside the display. This may not be surprising since the agent requires input in terms of written text using the keyboard. Not looking at the display means that we have no eye-data. This means that either the respondent blinks, eye-data is lost while the respondent looks at the screen due to a less optimal calibration or the person looks outside the display at the keyboard, at the task description or at the moderator. From the video of the respondents, it is clear that approximately 40% of the viewing time outside the display is spent on typing at the keyboard

Questionnaire Analysis

Four groups of eight persons each were recruited for the test sessions, two groups of novices and two groups of experienced users in order to test the validity of the COGITO prototype for each of these types of end-users. These groups may be treated individually or – if preferred for improving statistical considerations – added two by two as groups of 16 persons for each agent to be compared. As an example of the outcome concerning the criteria mentioned above, the result of the 'Impression – users feelings or emotions' has been presented in Figure 6 (for a full presentation of the evaluation see Andersen et al. [5]). The questions related to the impression of the agent are based on the agent being enjoyable or a bit awkward to use, and if the user would recommend use of the agent to colleagues. The groups of novices had rather negative feelings for both agents in this respect, probably because novices expect an agent - when being available - should act unimpeachable in all situations. The experienced users, however, are aware of the need of a period for maturing a new product, and in fact, the satisfaction among these users has increased from 44% for the state of the art agent (the sum of the percentages concerning "Exp. 1" for the columns "Satisfied" and "Very satisfied") to 61% for the COGITO agent (the sum of the percentages concerning "Exp. 2" for the columns "Satisfied" and "Very satisfied").

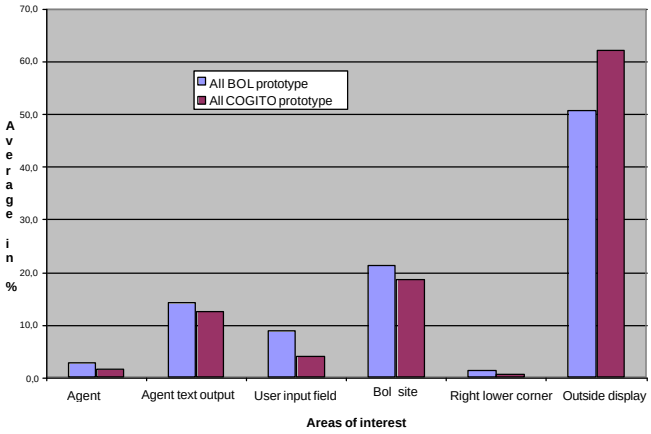


Fig. 5. The result of eye-tracking analysis for all respondents for the two agents. Numbers are averages in percentage.

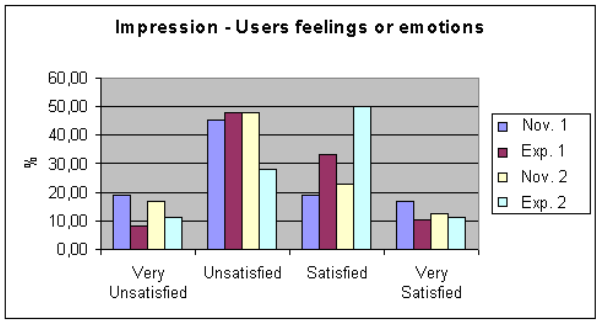


Fig. 6. User feelings or emotions. Numbers are shown in percentage of the total number of user ratings in each category. 'Nov. 1': novices from the baseline evaluation of the BOL state-of-the-art agent; 'Nov. 2': novices from the evaluation of the COGITO proactive agent, 'Exp. 1' and 'Exp. 2': similarly for the experienced users.

Conclusions

The COGITO agent has been evaluated for two groups of test persons, novices and experienced Internet users, to check if the agent is able to facilitate the interaction between the user and an e-commerce site exemplified in the COGITO project by the BOL site offering books, music, and gifts via the Internet. Even though the evaluation of the COGITO agent was preliminary, it is clear that the impressive amount of successful search results presented within the stop rule of two pages is due to the improved query, based partly on the expanded search criteria (due to the improved thesaurus) and partly on the search criteria presented directly by the user related to the improved conversation. From the questionnaire analysis the experienced users seem to appreciate more the developed improvements than the novices. Being more specific, for the topics selected for evaluating the satisfaction related to the impression of the agent this increased for the experienced users from 44% for the state of the art agent to 61% for the COGITO one. Likewise, for results not presented in this paper the satisfaction related to agent effectiveness increased from 38% to 63%. For the novices, however, the satisfaction related to learnability increased from 37% to 55%.

Acknowledgement

We would like to thank Osnabrück University for its hospitality in hosting the evaluation sessions, and especially Frank Ollermann for his help in preparing and being the moderator during the sessions. This work was partially supported by the project COGITO under contract IST-1999-13347.

References

[1] Abbattista, F., N. Fanizzi, S. Ferilli, P. Lops and G. Semeraro: *User Profiling in an Application of Electronic Commerce*. In F. Esposito (Ed.), *AI*IA 2001: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence 2175, 87-98, Springer: Berlin, 2001.

[2] Abbattista, F., O. Licchelli, P. Lops and G. Semeraro. *Usage Patterns Extractor*. Deliverable 4.3, COGITO IST-1999-13347, University of Bari, IT, 2001.

[3] Andersen, H.H.K., V. Andersen, C.B. Hansen, F. Abbattista, N. Fanizzi, F. Lioce, G. Semeraro and A. Thun. *Study of the end-user needs and behaviour*. Deliverable 7.1, COGITO IST-1999-13347, Risoe National Laboratory, DK, 2001.

[4] Andersen, V., H.H.K. Andersenand C.B. Hansen. *Establishing criteria for evaluating intelligent agents in E-commerce*, UM2001 workshop, Sonthofen, July 2001.

[5] Andersen V. and H.H.K. Andersen. *Evaluation of the COGITO system*. Deliverable 7.2, COGITO IST-1999-13347, Risoe National Laboratory, DK, 2002.

[6] Fenichel, C. H. (1981). *Online searching: Measures that discriminate among users with different types of experience*. Journal of the American Society for Information Science, 32, 23-32.

[7] Jansen, B.J., A. Spink and T. Saracevic, (2000). *Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web*. Information Processing & Management, 36, 207-227.

[8] Kobsa, A. *User Modeling: Recent Work, Prospects and Hazards*. In Schneider-Hufschmidt, M, Kuehme, T, and Malinowski, U (Eds.), *Adaptive User Interfaces: Principles and Practice North-Holland*, The Netherlands, 1993.

[9] L'Abbate M. and U. Thiel, (2001): *"Intelligent Product Information Search in E-Commerce: Retrieval Strategies for Virtual Shop Assistants"*. In Proceedings of E-work and E-Commerce Conference 2001, Venice, Italy. B. Stanford-Smith & E. Chiozza (Eds.), IOS Press 2001. ISBN 1 58603 205 4.

[10] Lesk, M. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann, San Francisco, CA, 1997.

[11] Lesk, M. *The Digital Library: What is it? Why should it be here? Source Book on Digital Libraries*. Technical Report TR 93-35, Virginia Tech, Dept. of Computer Science, Blacksburg, VA. Edited Volume, E.A. Fox (Ed.), 1993.

[12] Rasmussen, J., A.M. Pejtersen, and L.P. Goodstein. *Cognitive systems engineering*. New York: John Wiley & Sons, 1994 (Wiley series in systems engineering).

[13] Spink, A. and T. Saracevic, (1997). *Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching*. Journal of the American Society for Information Science, 48(8), 741-761.

[14] Tasso, C. and P. Omero, *Personalization of web content: e-commerce, i-access, e-government*, Franco Angeli Ed., Milano, 2002 (in Italian).

Digital Library Evaluation as Participative Inquiry

Ann Peterson Bishop
abishop@uiuc.edu

Bertram (Chip) Bruce
chip@uiuc.edu

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, IL 61820
217.333.3280 (V)
217.244.3302 (F)

Abstract

This paper outlines an approach to digital library evaluation that is participative and democratic. It has been developed especially to improve digital libraries whose users include marginalized society members, such as the poor and people of color. It differs from traditional approaches in that it is based on theories and methods drawn from participatory action research and inquiry-based learning.

Introduction

What particular issues arise in digital library research when intended users are on the wrong side of socioeconomic, information, and digital divides? How can digital library evaluation account more fully for the practices and consequences of use for marginalized members of society, such as the poor and people of color? The evaluation of information technology, generally, is too often characterized by features that are especially detrimental to those outside the social mainstream: it relies on external standards of expertise; treats users as subjects or objects of the evaluation; and pays only indirect attention to social outcomes associated with both evaluation practices and digital library use.

The Afya Project: Bringing Participatory Action Research and Collaborative Inquiry to Digital Library Evaluation

In the Afya (Swahili for "health") project, we are establishing a collaboration between local Black women--members of a grassroots group called SisterNet--and other community partners in building a collection of digital tools and resources that women will find congenial, usable, and useful in their efforts to nurture a healthy lifestyle. Our goal is to build capacity for creating and sharing health information across the social, cultural, economic and technology divides that separate Black women from health and information service providers. Ultimately, we are striving for social transformation on a community-wide basis. To support these aims, evaluation in the Afya project aims to accomplish three things: 1) Incorporate local knowledge held by marginalized groups; 2) Gain the participation of marginalized groups in evaluation activities early in the design and development process; and 3) Build capacity and achieve constructive social outcomes.

Our work in the Afya project represents a reconsideration of traditional approaches to information technology evaluation. To engage seriously with the social practice of disenfranchised users, we are incorporating ideas and techniques from two domains that are not often folded into technology evaluations. One domain is participatory action research (PAR), which claims social practice as its fundamental object of study and explicitly pursues an agenda focused on improving conditions for disenfranchised members of society (Reardon, 1998; Whyte, 1999). The other domain encompasses inquiry-based learning (Dewey, 1938). Here we find that framing evaluation as a collaborative "community inquiry" process helps in integrating the knowledge and views of diverse participants in the development of digital libraries, in a way is more democratic and in which everyone can learn from each other (Bruce and Bishop, 2002). In addition to foregrounding social responsibility, our participative inquiry approach to evaluation focuses on new ways of thinking about knowledge and how it is created and used by those with a stake in the process and outcomes of evaluation: social science researchers, digital library designers and managers, librarians, and marginalized members of society.

The purpose of this paper is to introduce the theoretical underpinnings of our community inquiry approach to digital library evaluation, presenting just a few examples of how this approach has been implemented in the Afya project. For more detailed reports on the Afya project's methodology, see: Bishop et al., 2000; Bishop et al., 2001a; Bishop et al., 2001b; Mehra et al., 2000.

Participatory evaluation (along with its near relative empowerment evaluation) extends participatory action research principles and methods into the domain of evaluation. It reflects concerns for both social justice and the improved utilization of evaluation results. Learning and empowerment are explicit goals of both the conduct and product of participatory evaluation, which is seen as a developmental process (Whitmore, 1998). There is a broader view of what constitutes legitimate knowledge, and how it is generated for and through action. Key participants are those with little power who will be affected by the evaluation. Their knowledge and how it can be engaged in improving social practice are central to evaluation.

Both the process and product of evaluation must respect multiple perspectives. In dealing directly with the question of racism in evaluation research, Patton (1999, p. 437) calls for evaluators to reflect more critically on their own knowledge bases and values: How does the lens of race, gender, or socioeconomic status shape the understanding and actions of researchers? What methods and measures capture fairly the experiences of people of color and the poor? When are evaluative judgments conditioned by personal politics?

Especially germane to the Afya project is the negotiation of whose knowledge is legitimate in developing and evaluating digital health resources. A cornerstone of inquiry-based learning is that it aims to respond to human needs by democratic and equitable processes. A successful "community of inquiry" (Peirce, 1868) is not one in which everyone is the same, but instead one that accommodates plurality. Clark argues (1994, p. 74) that we should focus on maintaining equitable relations first, and then consider collective tasks: “[This] renders the progress of expertise in a community secondary to a relational and epistemological practice of confronting differences so that its participants can come to understand how the beliefs and purposes of others can call their own into question.” Glassman (2001) notes that the “disturbed equilibrium” that occurs when the local knowledge held by diverse individuals comes into contact, and conflicts, is the necessary grounding for true learning and change in a democratic society.

In the Afya project, “use scenarios” that recognize the validity of knowledge arising from the values, goals, and experiences of SisterNet women help us achieve a more socially just equilibrium in digital library design and evaluation (Bishop et al., 2001b). An initial set of scenarios related to the use of networked health information services were developed from data gathered in four focus groups. Three focus groups were conducted with SisterNet women. In addition, one focus group was conducted with participants who were community health care and health information providers. About six to eight people participated in each group. SisterNet focus group participants discussed the following:

- Their vision of a healthy Black woman;
- Important health concerns or situations they recently experienced;
- Where and how they typically get and use health information;
- Barriers they experience and what works well in using health resources;
- Their use of computers; and
- Actions that they or other Afya project participants could take to improve health information resources and services.

Local health care and health information service staff addressed similar issues, but from their own perspectives as providers. These topics were integral to developing use scenarios for a web-based community health information service for African American women because they holistically bring together essential needs, goals, expectations,

and practices related to the use of health services, information, and technology. For example, here is one scenario contributed by a SisterNet woman:

"My daughter had mono [...] and she lost like twenty pounds. [...] My mother is a nurse practitioner and the first thing she said was that my daughter needed to go to the hospital and be taken care of. So I did take her to the doctor and the first thing they said was "well, we ..." I was hitting a brick wall [and] had to take her to the emergency room to get her the care she needs. She was severely dehydrated and was going downhill. You know if you look at the child and that child is being dragged in by the parent because the child cannot walk... How severe does one have to be before they can get the care they need? Before you can get someone to listen to you?"

Another aspect of reframing evaluation is its use as a concept and technique throughout the lifecycle of system design and implementation. In Afya, evaluation is cast as the fundamental task and responsibility of the SisterNet women serving as our community action researchers, an activity that began in our needs assessment discussion groups, where women presented their assessment of community healthcare and health information provision. In subsequent situated evaluation workshops, SisterNet women learned how to find health information on the web. They then assessed health websites according to their relevance to the scenarios generated earlier. They also commented on the sites' ease of use and appeal to Black women. In these workshops, SisterNet women were not treated as subjects undergoing a laboratory-style usability experiment. The workshops were held in a pleasant and informal setting, their children were welcome to come along, and refreshments were served. The SisterNet women were treated as the evaluation experts (while faculty members and students recorded their evaluative comments). In addition, they received several direct benefits, including print guides to health related websites and online discussion groups, along with "door prizes" (popular books on Black women's health).

We believe that the processes of creating, using, and evaluating the tools and resources of digital libraries should exemplify the social participation and capacity-building that both PAR and inquiry-based learning exemplify. Digital library users are the developers, through their creation of the site content, their contributions to the interface, and their evaluations, often simply by discussion within the inquiry community of its usefulness, and reports of what works and what does not in the context of their own settings of use. Workshops, conducted in a variety of settings, are the cornerstone of evaluation through participative inquiry.

In another Afya project activity, a small group of SisterNet women, university faculty and students, and librarians gathered to brainstorm about a workshop for their Spiritual Health Conference that would simultaneously promote learning about both computers and spiritual health. As always, SisterNet was keen on developing a workshop activity that involved Black women taking direct action related to improving their lives, not just passively absorbing relevant information. At the pre-conference brainstorming session, the women present decided to investigate use of the Inquiry Page Web site (http://www.inquiry.uiuc.edu) for their spiritual health and technology

workshop. The tossed around some ideas and ended up creating an Inquiry Unit in the form of an online template that SisterNet women could "spin off" to create their personal spiritual health plans. Librarians, students, and SisterNet women all suggested Web sites related to Black women's spiritual health that could be included in the template, and together they crafted the template's instructions for creating a personal health plan. Arlene was one of the women who participated in the workshop. Her spiritual health plan can be viewed at: http://www.inquiry.uiuc.edu/bin/update_unit.cgi?command=select&xmlfile=u10905.xml

The spiritual health and technology workshop demonstrated how diverse participants in community inquiry can learn from each other in the course of digital library evaluation. To give one simple example, SisterNet women who created spiritual health Inquiry Units learned how to find spiritual health information on the web. Inquiry Page developers, in turn, learned how to improve the Inquiry Unit web form so that it would better suit members of the general public. When SisterNet women noted that the emphasis on grade levels and school subjects in the Inquiry Unit form was not appropriate for their community-action project, the developers found a way to change the presentation of the Units so that grade levels and school subjects were not so prominent. We are now working on a "distributed Inquiry Page module" which customizes the Inquiry Page for the SisterNet website. At a subsequent Afya workshop, participants explored using their prototype customized Inquiry Page to develop other types of online health-related resources that would bring together ideas, information, and experiences from the perspectives of different community members. They came up with an idea for creating simplified Inquiry Units that all SisterNet women could easily use to share their health stories, questions, and tips in a kind of "SisterNet Scrapbook" on their Web site. SisterNet women--many of whom are low-income community residents with little computing experience--are proving that people on the wrong side of the Digital Divide can become active creators and contributors in the development and evaluation of digital libraries.

Our work in the Afya project is helping us learn about digital library evaluation as a socially grounded activity that is based in professional social responsibility, attends to details of social practice, and can lead to positive social consequences. An evaluation approach that closes the distance among digital library stakeholders by framing their work as collaborative inquiry is crucial for developing valuable and equitable online services in a world scarred by social, economic, and digital divides.

References

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.org/ issues/issue5_6/bishop/index.html]

Bishop, A. P., Bazzell, I., Mehra, B., & Smith, C. (2001a). Afya: Social and digital technologies that reach across the Digital Divide, *First Monday*, *6*(4). [http://www.firstmonday.org/issues/issue6_4/]

Bishop, A. P., Mehra, B., Bazzell, I., & Smith, C. (2001b). Use scenarios in the design and evaluation of networked information systems: An example from community health. In McClure, C., and Bertot, J. C. (Eds.). *Evaluating Networked Information Services: Techniques, Policy, and Issue*s, pp. 45-66 Medford, NJ: Information Today.

Bruce, B. C., & Bishop, A. P. (2002, May). Using the web to support inquiry-based literacy development. *Journal of Adolescent and Adult Literacy*, *45*(8).
[http://www.reading.org/publications/jaal/index.html]

Clark, G. (1994). Rescuing the discourse of community. *College Composition and Communication*, *45*(1), 61–74.

Dewey, J. (1938). *Experience and education*. New York: Macmillan.

Glassman, M. (2001). Dewey and Vygotsky: Society, experience, and inquiry in educational practice. *Educational Researcher*, *30*(4), 3-14.

Mehra, B., Bishop, A.P., & Bazzell, I. (2000). The role of use scenarios in developing a community health information system. *Bulletin of the American Society for Information Science*, *26*(4), 10-12.

Patton, M.Q. (1999). Some framing questions about racism and evaluation: Thoughts stimulated by Professor Stanfield's "Slipping through the Front Door." *The American Journal of* Evaluation, 20(30), 437-444.

Peirce, C. S. (1868). Some consequences of four incapacities claimed for man. *Journal of Speculative Philosophy*, *2*, 140-157.

Reardon, K. M. (1998). Participatory action research as service learning. In R. A. Rhoads and J. P. F. Howard, eds., *Academic service learning: A pedagogy of action and reflection* (pp. 57-64). San Francisco: Jossey-Bass.

Whitmore, E. (ed.). (1998). *Understanding and practicing participatory evaluation.* San Francisco, Jossey-Bass.

Whyte, W. F. (ed.). (1999). *Participatory action research.* Newbury Park, CA: Sage.

Evaluating a Digital Library for Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype (ADEPT)

Presentation for the Fourth DELOS/NSF Workshop on Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics, 6-7 June 2002, MTA SZTAKI, Budapest, Hungary

Christine L. Borgman
University of California, Los Angeles

Co-Investigators, Education & Evaluation team:
Anne Gilliland-Swetland, Gregory Leazer, UCLA; Richard Mayer, UCSB
Student Researchers: Jason Finley, Rich Gazan, Kelli Millwood, Laura Smart, Annie Zeidman (UCLA); Tricia Mautone, Rachel Nilsson, UCSB
Research funded by U.S. National Science Foundation, Digital Libraries Initiative Phase 2, grant no. IIS-9817432, Terence R. Smith, University of California, Santa Barbara, Principal Investigator.

While many forms of digital libraries exist, each system does so in a specific context. In an educational context, digital libraries can support the delivery of content to classrooms, labs, dorms, offices, and homes. They can facilitate instruction. They can provide primary and secondary resources. Digital libraries can provide services to instructors such as constructing lectures and laboratories. Students can explore and manipulate digital resources. In a university context, digital libraries function within an infrastructure for teaching and research. Digital libraries can supplement library resources and services, and they can supplement instructional development resources and services.

We report on the first three years of a five-year educational evaluation study (1999-2004) of a geographic digital library in undergraduate education at the University of California, Los Angeles (UCLA) and the University of California, Santa Barbara (UCSB) [1], [2], [3], [4]. This study is part of the Alexandria Digital Earth ProtoType (ADEPT) project; The ADEPT web sites at UCLA (http://is.gseis.ucla.edu/adept/) and UCSB (http://www.alexandria.ucsb.edu/adept/) will provide links to continuing reports of ADEPT research.

ADEPT seeks to provide instructors, teaching assistants, and students with the means to discover, manipulate, and display geographical content. Geography is a particularly fruitful area for studying the role of digital libraries in scientific thinking. It is a scholarly discipline that relies on a rich array of primary data sources such as maps, satellite observations, remote sensing, and physical observations. Geography also is a discipline that studies dynamic processes using models and simulations.

Geographic information systems are widely available and geographic learning is relatively well documented. Five skill sets for scientific thinking in geography have been

identified, all of which can be aided by digital libraries: (1) asking geographic questions, (2) acquiring geographic information, (3) organizing geographic information, (4) analyzing geographic information, and (5) answering geographic questions [5]. Similarly, the U.S. National Science Education Standards [6] emphasizes "inquiry into authentic questions generated from student experiences." Although we focus on promoting geographic thinking in our project, the same kinds of skills that support geographic thinking also apply to other disciplines. The five skills can be used in inductive (or data-driven) reasoning, such as looking for trends in data that lead to a theory, or in deductive (or theory-driven) reasoning, such as testing two competing theories through dynamic modeling.

Our evaluation is driven by a series of desired outcomes of the ADEPT project. If we are successful, the use of ADEPT will encourage instructors to incorporate primary sources in their teaching and to use digital library services to enrich instruction. As a consequence, students using ADEPT will become more active learners and will learn to "think like scientists" by locating relevant information, balancing evidence, synthesizing knowledge, testing hypotheses, and by developing their own conclusions.

ADEPT is a large, interdisciplinary project being conducted at several universities. For the purposes of this presentation, ADEPT can be viewed as a set of two complementary teams. The Implementation team (which has several component teams), based at the University of California, Santa Barbara (UCSB), focuses on how ADEPT is constructed. The Education and Evaluation team (the authors of this paper, at UCLA and UCSB) focus on how ADEPT is used. We are conducting formative evaluation, which involves a series of needs assessments as input to the design process, and iterative design, which involves continual evaluation of ADEPT usability in learning situations. We provide feedback to the design process based on our evaluations. At the end of the project, we will conduct summative evaluation.

Our evaluation efforts require a variety of quantitative and qualitative research methods. These include (a) classroom-based studies in which we interview faculty (academic staff), teaching assistants, and students, observe lectures, and gather data on student performance and demographics; (b) laboratory-based studies, such as assessment of mental models and scientific thinking processes; and (c) office-based studies, in which we interview faculty about their information seeking in support of teaching and research. We envision that a typical use of ADEPT for a topic such as river networks would proceed as follows:

Instructor
–Prepares class lecture with ADEPT
–Discovers relevant geographic objects
–Constructs metadata to describe objects for personal and shared use
–Annotates objects as necessary
–Integrates objects into Iscapes (personal digital libraries)
–presents lecture to students using ADEPT

Teaching assistants
–review selected topics in lab sessions using ADEPT
Students
–use ADEPT for lab exercises and to study for exams

The talk will present examples from the current ADEPT prototype.

Our evaluation plan has evolved over the course of the project. Some highlights are these:

Evaluation, Year 1 (1999-2000)

In the first year of the ADEPT project, we studied 4 sections of the course, "Geography of the Physical Environment." Three of these sections were taught at UCLA and one at UCSB (which is a much smaller campus). Each section was taught by a different instructor, and 2 instructors were involved in the UCSB course, for a total of 5 instructors to study the first year. During that period of time, we developed our data collection instruments, observed classrooms, gathered baseline data on student demographics, and conducted interviews with faculty, students, and instructional development staff.

We were conducting baseline data in parallel with the development effort at UCSB, and thus little software was available for testing. As a means to test our instruments and to get some initial data, we mocked up a prototype in MS Powerpoint that incorporated slides, scanned images, sound, and QuickTime movies. These resources were gathered from personal collections, text books, web sites, and other sources. The course topics that we mocked up were on fluvial processes and on hydrology.

Our initial data indicated instructors' approaches to teaching the same course vary widely. While the Alexandria Digital Library (constructed as part of the earlier NSF DLI-1 project) contains a rich set of geospatial resources, instructors still wished to integrate additional multimedia resources into ADEPT collections. The instructors we studied were heavy users of information technology for their research, but they made minimal use of IT for teaching. In gathering and selecting resources for the mockups, we found that display, layout, and other presentation and navigation features are important considerations.

Evaluation, Year 2 (2000-2001)

In the second year of the project, we focused on observing classrooms to assess variation in teaching styles, concept integration, and use of technology. We gathered data to draw concept maps of individual instructors' lectures, developed use scenarios to inform design, and assessed metadata requirements for educational uses. We also evaluated the Year 1 mockups with additional instructors.

Current activities, Year 3 (2001-2002)

In the current academic year, the education and evaluation team is continuing the needs assessments, refining the use scenarios, and refining the course topographies and concept maps. The teams are jointly working on graphical representations of course content and concept linkage.

We have begun several new efforts this year. A study is underway to assess the roles of teaching assistants in teaching geographic concepts, in teaching scientific thinking, and in collaborating with instructors. Another pilot study is being conducted to study the information-seeking activities of faculty as geography instructors and as geography researchers.

The Implementation Team is in the final stages of developing a working prototype, which our team will evaluate with instructors and teaching assistants, based on the scenarios developed in year 2.

Evaluation plan, Year 4 (2002-03)
Next year we plan to implement the ADEPT prototypes at UCSB and UCLA and to undertake a series of evaluations of the prototype implementation. These include classroom observations, laboratory assessments of scientific thinking, and interviews with instructors, teaching assistants, and students. We will assess learning outcomes in the courses, will observe students in scientific problem solving, and will assess student performance on the five categories of geographical knowledge outlined above (asking geographic questions, acquiring geographic information, organizing geographic information, analyzing geographic information, and answering geographic questions).

Evaluation plan, Year 5 (2003-2004)
In the last year of the project we plan to conduct summative evaluation of teaching and learning. This includes the full implementation of ADEPT at UCSB and UCLA, laboratory assessments of scientific thinking, and follow up on students who learned geography with ADEPT. We will devote efforts to disseminating resources and results and hope to provide distributed access to ADEPT collections and services.

Summary
Our experience with the ADEPT project indicates that evaluation of digital libraries must be context-specific. It is important to focus on real uses of specific types of content, on real users who have a variety of uses. Our evaluation has complementary goals: Systems that facilitate desired outcomes (e.g., scientific thinking) and systems that will be adopted and used. We feel that evaluation should be holistic, by gathering data on users and uses in situ, and it should be iterative, providing input to the design process.

References

1. Leazer, G.H., Gilliland-Swetland, A.J., Borgman, C.L.: 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, (2000) 248-249.

2. Leazer, G.H., Gilliland-Swetland, A.J., Borgman, C.L., Mayer, R.: Classroom Evaluation of the Alexandria Digital Earth Prototype (ADEPT). Proceedings of the American Society for Information Science Annual Meeting (2000) 334-340.

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

4. Gilliland-Swetland, A.J., Leazer, G.H.: Iscapes: Digital Library Environments to Promote Scientific Thinking by Undergraduates in Geography. Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries (2001).

5. Geography Education Standards Project: Geography for Life: National Geography Standards. National Geographic Society, Washington, DC (1994).

6. National Research Council: National Science Education Standards. National Academy Press, Washington, DC (1996).

Usability Evaluation in the Context of Digital Video Archives

Joanne Evans[1], Andrew O'Dwyer[2], Stephan Schneider[3]

1 Introduction

This contribution reports briefly the usability evaluation that has been performed in the EC funded AMICITIA project (IST1999-20215) and in the PRIMAVERA project (IST1999-20408).
Both projects are developing novel features and extensions to large video archives for broadcasters and national institutions. The developments follow a "user-driven" approach which means that the development is in close relation to user needs and evaluation. After the initial implementation of the new software modules an intensive usability testing is applied to refine and improve the system.

2 Standards

„Usability" can be considered as part of „software quality". Software quality is defined e.g. in the **ISO/IEC 9126** standard. The ISO/IEC 9126 Software Quality Model defines six features for ensuring software quality:

1. Functionality
2. Reliability
3. Usability
4. Efficiency
5. Maintainability
6. Portability

Our specific interest is in the functionality, usability and efficiency of the software.
We need 'usability' in our projects in order to achieve system acceptability. Acceptability is most usually thought of as consisting of a number of different factors, in particular the *usefulness* and *ease of use*[4] of the system ([1]). A consequence of this is satisfactory use of the system, which is a digital video archive, and thus acceptance. However, it is important to note that satisfaction is not always a consequence of usability, but ought to be considered a factor in its own right. ([2], [5]). As such the general theoretical framework adopted for assessing system acceptability is:

1. Utility
2. Usability
3. Satisfaction

Another ISO standard, the **ISO 924-11** Guidance on Usability, defines usability in more detail:

Usability	The extent to which a product can achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.
Effectiveness	The accuracy and completeness with which users achieve specified goals.
Efficiency	The resources expended in relation to the accuracy and completeness with which users achieve goals.
Satisfaction	The comfort and acceptability of use.

For pragmatic reasons, the distinctions between effectiveness and efficiency are grouped as a single feature of *utility*. However, in any assessment of utility both effectiveness and efficiency has been measured.

3 The Dimensions of Usability

As discussed before, three aspects of usability will be assessed:

1. Utility
2. Usability
3. Satisfaction

This is in overall conformance with ISO standards, but has the simplification of using the category Utility to represent the ISO standard categories of Effectiveness and Efficiency. All these terms must of course bee seen in the context of digital archives or libraries.

---

[1] BBC, Information & Archives, London, UK
[2] BBC, Information & Archives, London, UK
[3] tecmath AG, Content Management Systems Division, Kaiserslautern, DE
[4] *Usefulness* and *ease of use* are considered synonyms for *utility* and *usability*.

3.1 Utility

Utility will be assessed using performance measures on task, such as accuracy, completion time, key presses and errors made. For the purposes of our evaluations effectiveness and efficiency will be a feature of utility.

Utility can be measured objectively in terms of time needed to complete a given task. For all our evaluations, there should be one or more standard tasks that can be performed on base systems and on newly developed software, to support gathering of objective utility statistics. This task or tasks needs to be performed on differing sets of media or data, in order to prevent users having to exactly repeat tasks, and in order to minimise learning and order-of-performance effects.

Concerning the digital archive under evaluation these tasks must be typical tasks that users have to perform when they are working with the archive. In our first evaluation the focus was on retrieval and browsing of digital video items. Subjective information about a user's attitude toward utility can be obtained from a questionnaire or an interview. There should be a formal interview (de-brief) as a part of all our usability evaluations.

3.2 Usability

Usability will be measured using an adaptation of the heuristic approach proposed by Nielson ([4]):

Nielson: Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
Match between system and the real world	The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards	Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
Error prevention	Even better than good error messages is a careful design that prevents a problem from occurring in the first place.
Recognition rather than recall	Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
Flexibility and efficiency of use	Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
Aesthetic and minimalist design	Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Help users recognise, diagnose, and recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

In our evaluation the criteria listed above have been applied to a web based retrieval interface and a browsing player for video items. The tailoring of the evaluation and the practical implementation are described below.

3.3 Satisfaction

Satisfaction will be assessed using specifically developed questionnaires based on adaptations of existing subjective response scales. There are many features that determine satisfaction; examples are:

1. Learnability
2. Efficiency
3. Usefulness
4. Aesthetics

4 Methodologies and Assessment Tools

A number of different methodologies will be employed that hinge around a common set of assessment tools. The key difference between methodologies is that of field based observation, possibly employing structured interviews and laboratory based task led studies of acceptability.

4.1 User selection and Briefing

All our testing is to follow very simple ethical guidelines. All users asked to take part in user testing are to be briefed (in writing or verbally). The following information is key.

1. Anonymity will be preserved, particularly any data held by the testers.
2. The user can chose to leave the test at any time.
3. The user is free to ask questions, though this would be more convenient at the end of the study.

4.2 Demographic Data

A core set of demographic data is required in order to determine the user profile and check that system acceptability factors are not the direct result of variables such as sex and age. Tracking these variable and cross checking during analysis for their effects is vital ([6], [7]).

The relative levels of experience and frequency of use must be measured. Furthermore, it is worthwhile to ensure that measurements towards attitudes to computers and technology can provide interesting explanation of acceptability effects.

4.3 Field Observation

Testing a product in the actual context in which it will be used (as opposed to laboratory testing, or testing in the development environment) can be useful for highlighting a wide range of context characteristics and interaction problems.

4.4 Context Analysis

Context analysis is used both to characterise user profiles and identify user tasks and environmental factors.

1. The operating environment
2. The users of the product
3. The tasks they carry out
4. The social and organisational environment

4.5 Task Analysis

Through observation, structured interviews and questioning, a detailed understanding of the users tasks and goals are identified.

1. The job they do
2. Their goals
3. Their Domain knowledge
4. Their Procedural Knowledge

4.6 Evaluation Procedure

The user-testing phase will follow a standard sequence and lead to a final usability report. The standard sequence for Usability evaluation is:

Select the users > Design the test > Set up the test > Run the test > Analyse the results > Write the report

5 Implementing the Evaluation

A total of twenty staff from within the BBC participated in the usability evaluations. They were broken down into the four user profile groups prior to the evaluation. Testing was carried out at the BBC Information & Archives offices in London over a number of weeks.

5.1 Target Application

The application, a digital video archive, is installed in a number of partner sites and varies in design and complexity of use. Content is digitised through a process called ingestion. Once digitised, keyframes are selected that best represent the content. It is catalogued using keywords and additional information such as broadcast date and spool number are added.

Search and retrieval of content is achieved using the retrieval interface. Content can be displayed either in the form of keyframes that represent the item, or displayed as text information.

Keyframes give a pictorial representation of the events within an item and are supported by cataloguing. This gives the viewer a quick overview of what content is available. A colour-coded symbol indicates if any restrictions on reuse or copyright apply.

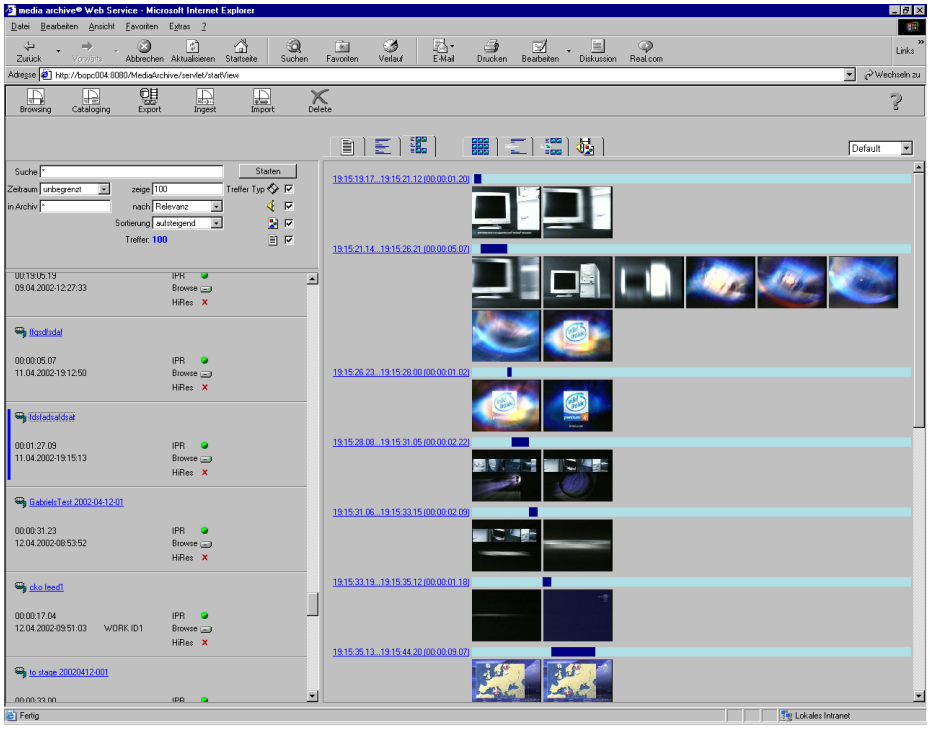
Clicking on an icon or any key frame opens a video browser window and the item is played in low resolution. The data rate is 1.5 Mbit/s, (MPEG1). Using on screen controls it is possible to stop, start, or replay an item from the video browser.

The evaluation has been focused on three components of the whole application.

1. The Retrieval Interface – used to search and display results
2. The Browsing Player - plays video content
3. The Browsing Editor – used to create edit decision lists

The above components have been chosen because they are the most widely used in a digital archive installation. Additionally the users that use the retrieval interface form the most heterogeneous group ranging from naïve users up to experts.

The figure below shows a screenshot of the Retrieval Interface that has been evaluated:



The BBC has over fifty hours of material on the application that is being used for this evaluation. The content is a mixture of whole programmes and rushes from the Holiday Programme.

The cataloguing for this material is in the form of keywords. These keywords relate to title, location, subject matter and actuality within the item. The application is started through Internet Explorer and can be operated on a number of clients simultaneously.

5.2 Evaluation Method

As stated earlier the objective is to evaluate the application through a series of task led activities. The will probe the application for possible usability problems. These tasks will reflect the use of the application in a working environment and the tasks will represent the typical use of the application.

During testing two observers will note the subject's interaction and gather usability information. Following the tests subjects are interviewed about their experiences and asked to fill out a post-test questionnaire. The interview and questionnaire are designed to gather more qualitative assessment on features such as likes, dislikes and usefulness.

5.3 Subjects

In usability testing participants are referred to as subjects and we will conform to this convention throughout this paper. The subjects for this specific usability tests have been drawn from the staff of BBC Information & Archives. Subjects recruited have filled out a profile user questionnaire to establish their knowledge, experience, qualification and attitudes to computers and technology. All subjects are advised that the information and data collected over the course of the usability test is confidential.
It is expected that people with varying skills and abilities, as well as cultural and language differences will use the system. Field-testing can identify cultural usability problems not readily evident in laboratory tests, such as environmental and interaction factors.

5.4 The Test Room

For the evaluation we've used a room within the BBC Information & Archives offices. It is a normal office environment with desk chairs and computers. The subject sat at a desk with a computer. Additional seating was provided in the room for the *test monitor* and *two observers*.
The *observers* are concerned with observing interactions between the subject and the browser. They are responsible for capturing data throughout the test such as task time, error rate, verbal responses and heuristics.
The *subject* is asked to carry out a number of set tasks given to them by the test monitor. They are encouraged to talk through what they are doing during the test to help the observers gather information. Following the test the subject attends a post-test interview to provide a qualitative assessment on the use of the application..
The *test monitor* is the contact with the subject during the test. Their role is to initiate a set of tasks and steer the subject through the test.
A number of *scripts* and *documents* are used before, during and after the usability test session. These are checklists and forms used during the test. Examples of these are available in the handout.

6 Test design

Profile User Groups Questionnaire
Prior to attending the test, subjects are sent a profile user group questionnaire. From this subjects are selected into appropriate user groups dependent on factors such as education, experience and knowledge. For the purpose of our evaluation the profiled user groups have been determined into four sets.

Naive Occasional	Defined as, non-regular users of media catalogues and databases.
Naive Regular	Regular access to archive media, but not having librarianship training or qualification.
Professional Occasional	Managers with librarianship qualification but little experience of media catalogues.
Professional Regular	Professional information researchers, many with formal librarianship qualifications.

Task List

For these tests a number of tasks have been designed to bring into use functions associated with using the application. The aim is to probe the user interface to uncover potential usability problems. The set of tasks has been ordered in a way to expose the subject to the functions and features of the application and assess their responses. For the complete task list please see the handout.
The test has been broken down into three parts. The subject carries out a number of tasks initiated by the test monitor. After a break the subject is interviewed and then completes a post-test questionnaire (see handout).
The observers use an *observation form* during the test and collect information about their observations during the test. This is a simple form where the task number is entered along with the observer's notes. An example of this is in the handout.
As a baseline we have observed how long it takes an experienced and moderate user to carry out the tasks. We have allowed for *30 seconds per task* as a practical average time during testing.

A *pilot test* was run with staff not involved in the testing to provide training for the observers and test monitor. It was also used to ensure that the tasks themselves and the order in which they were presented to the subject were consistent with everyday use of the application.

7 Results

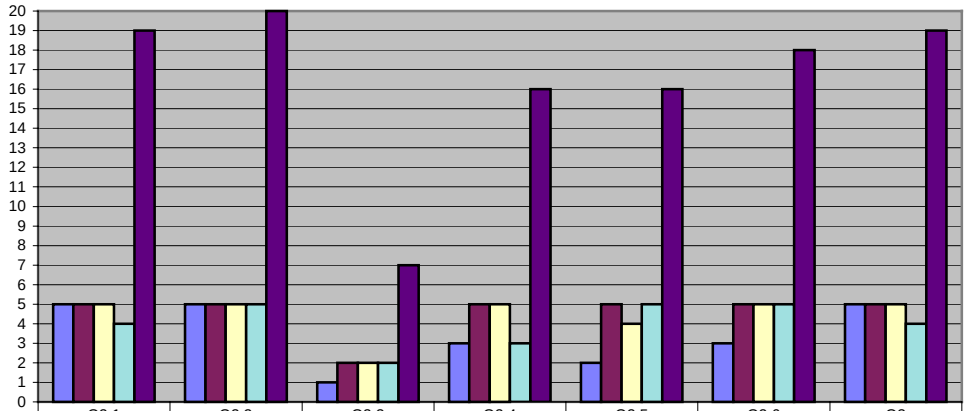
The data gathered from the usability tests have been entered into spreadsheets for tabulating and summarising the data. These results are displayed as a whole and broken down for more detail into specific profile user groups.
The more subjective data gathered from post-test interviews have been assessed in reference to the list of heuristics appearing earlier in this report. Satisfaction has been measured through the post-test interview and questionnaire.
The results and findings have been arranged into their respective sections of Utility for performance measures, Usability for heuristic (human interaction) results and Satisfaction for subjective reporting. A number of charts are provided with task achievement data and results from the subjective rating scale in the post-test interview.
From these section the results for Usability have been chosen as a brief example below.

Utility and Task Results

The data collected has been entered into charts to make the comparison between tasks and performance results in each user group easily comparable. The three charts have been broken down into the corresponding sections in the task list. The chart below shows the performance measure of the task "Simple Search":

Simple Search

(users where marked 1 for successfully completing the task and 0 for failing to complete the task)



	Q2.1	Q2.2	Q2.3	Q2.4	Q2.5	Q2.6	Q3
Naïve occasional	5	5	1	3	2	3	5
Naïve regular	5	5	2	5	5	5	5
Pro occasional	5	5	2	5	4	5	5
Pro regular	4	5	2	3	5	5	4
Total	19	20	7	16	16	18	19

Task number

Usability Results

Usability results have been collected from observation and post-test interviews. The two observers present during testing noted the subject's interaction with the application and their verbal responses. Subjects were encouraged before and during the tests to provide a dialogue as they carried out the tasks. This method of collecting information about the application and interaction issues provided a lot of useful data on usability factors.
During testing the observers matched their observations and dialogue from the subjects and matched them against an adaptation of the heuristic approach proposed by Nielsen and Molich ([8]).
Their heuristic approach contains a number of common usability principles that are generic in computer application evaluations. These principles address a number of factors:

1. Flexibility and efficiency of use
2. User control and freedom
3. Help users recognise, diagnose, and recover from errors

4. Help and documentation
5. Consistency and standards
6. Error prevention
7. Recognition rather than recall
8. Match between system and the real world
9. Aesthetic and minimalist design.

Using these usability principles as a guide, the test observers noted them into the Observation Report Form. The usability principles were noted alongside what the subject was doing and the task they were carrying out. This made for easy comparison and helped greatly in compiling the information.

Sample usability results

A lot of data has been collected during the test. Only a very small fraction can be presented here. The brief examples below are from Subject No: UP 213.

Observation report Form:

Task No	Usability Factor	Comments & Notes
2	**System visibility and consistency and standards.** **Match between system and the real world**	2.1 Confusion on why an asterisk was present in search box. 2.2 What does "segmentlist" mean on icon label – task failed
3	**Consistency and standards**	3. On trying to return to search screen – Subject noted that two icons have the same symbol of (?) – This caused confusion, used back button – task achieved
4	**Help and documentation** **Match between system and the real world**	4.1 (combined search) looked for help screen – found, but in German. Task achieved after confusion over icon some terminology. 4.6 Play Video Clip – "What does browse object mean on icon"
…		

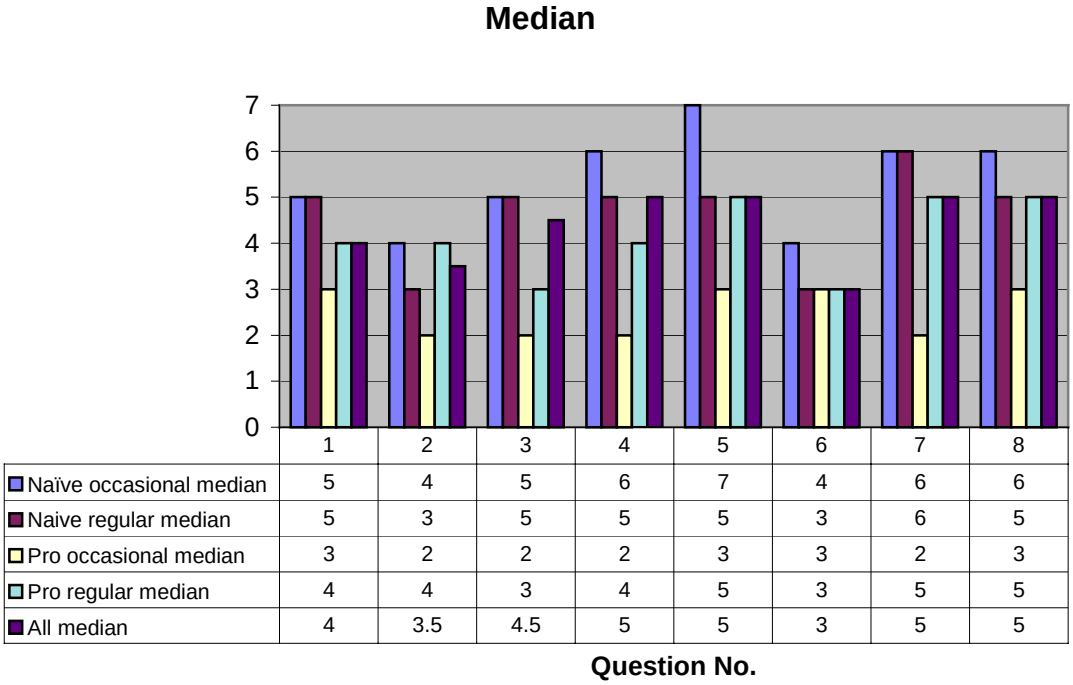
Usability Observations and Comments Table:

Usability Principles	Observations and comments
Flexibility and efficiency of use	- Liked the ability to see content quickly. - Liked the option of displaying just text. - Navigation could be easier.
Help users recognise, diagnose and recover from errors	- Would like the system to prevent them from logging out by mistake.
Help and documentation	- Subject looked for tutorial - Found help screen (in German)
…	

7.3 Satisfaction Results

Satisfaction has been assessed through post-test interviews and a subjective response scale questionnaire. Subjects were asked about their overall impression on using the application. They were asked about what they thought of the application and any likes and dislikes. The questionnaire was in a form of rating scale of 1 - 7 and the results from this have been tabulated for reference.

The chart below shows the user satisfaction assessment. It should be read in conjunction with the Post-test questionnaire.



7 Conclusion

The evaluation of the application highlighted usability issues that were consistent across all user groups. These related mainly to the icons, terminology and access to functions such as the browsing editor and change entry box.

The subjects enjoyed using the application and could see the benefits of it in the context of access to archive material. The learning curve in searching for content was good, with an increase in task achievement from simple, to advance search. User groups found varying satisfaction in using the application. The ability to search and view video material quickly was well received.

From an overall point of view the evaluations described has provided a sound basis for amendments and corrections in the next version of the prototype.

References

[1] Davis, F.D., User acceptance of information technology: system characteristics, user perceptions and behavioural impacts. International Journal of Man-Machine Studies, 38, pp. 475-487, 1993

[2] Winter, G.Q., Understanding the obstacles and opportunities for popular interactive services. Proceedings of the International Broadcasting Conference, Amsterdam, 2000

[3] Bevan, N., Industry Standard Usability Tests, INTERACT '99, 7th IFIP TC13 Conference on Human Computer Interaction, Edinburgh, 1999

[4] Nielson, J., Usability Engineering. Academic Press Professional, 1993

[5] Hassenzahl, M., Platz, A., Burmester, M., Lehner, K., Hedonic and Ergonomic Aspects Determine a Software's Appeal. Proceedings of CHI 2000, The Hague, Netherlands, March 2000.

[6] A Practical Guide to Usability Testing (ISBN 1-84150-020-8)

[7] Handbook of Usability Testing (ISBN 0-471-59403-2)

[8] Nielsen and Molich, Usability, 1990

Creating a Multilingual Test-Bed for Cross-Language System Evaluation

Carol Peters[1] and Martin Braschler[2]
[1]IEI-CNR, Area di Ricerca CNR, 56124 Pisa, Italy
carol@iei.pi.cnr.it
[2]Eurospider Information Technology, Zürich, Switzerland
martin.braschler@eurospider.com

Abstract

The aim of the Cross-Language Evaluation Forum (CLEF) is to develop and maintain an infrastructure for the evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes. The paper[1] discusses the criteria that have been followed in the creation of the multilingual test-bed and provides information on the evaluation methodology adopted and the actual contents of the test collection.

1 Introduction

The Cross-Language Evaluation Forum (CLEF) aims at promoting cross-language information retrieval (CLIR) system development by providing the research community with an infrastructure for:

- testing and evaluation of information retrieval systems operating in both monolingual and cross-language contexts
- objective comparison of different systems and approaches
- exchange of experiences and know-how between R&D groups working in the field.

The activity is conducted through a series of annual system evaluation campaigns. We are now in the process of organizing the third of these campaigns: CLEF 2002[2].

The design of the tasks offered by CLEF is studied to meet the needs of developers working mainly with European languages. However, strong links have also been forged with the other two major CLIR system evaluation activities: the Text REtrieval Conferences (TREC), run by the US National Institute of Standards and Technology, which are currently focusing on English/French to Arabic retrieval (Gey & Oard, 2001) and the NACSIS Test Collection for Information Retrieval (NTCIR), sponsored by the National Institute for Informatics of Tokyo, which offers cross-language system evaluation for Asian languages (see Kando et al, 2001). The three initiatives (US, Asian and European) aim at creating a network of complementary activities in the cross-language system evaluation area.

In this paper, we describe the criteria that have been followed in the creation of the multilingual test-bed and provide information on the evaluation methodology and the actual contents of the test collection. The paper is organised as follows. Sections 2 and 3 outline the methodology adopted by CLEF and list the main tasks currently offered. Section 4 describes the contents of the actual test collections and discusses the criteria followed in their creation, Section 5 discusses the techniques used for the results analysis, and Section 6 raises the important issue of the preparation of reusable test collections. Finally, we briefly mention our intentions for the future.

2 Methodology

So far, CLEF has focussed on measuring system performance in terms of successful retrieval of relevant textual documents. Following the model used in TREC, CLEF uses a comparative evaluation approach and has adopted the well-known Cranfield methodology (Cleverdon, 1997): performance measures are calculated based on a test

[1] This paper is based to a large extent on an article presented at LREC2002, Las Palmas, Canary Islands, Spain, and published in the LREC2002 Proceedings.

[2] CLEF 2000 and 2001 were sponsored by the DELOS Network of Excellence for Digital Libraries; from October 2001, CLEF is funded by the European Commission under the IST programme (IST-2000-31002). The consortium members are: IEI-CNR, Pisa; IZ Sozialwissenschaften, Bonn; ELRA/ELDA, Paris; Eurospider, Zurich; UNED, Madrid; NIST, USA.

collection, sample queries and relevance assessments for these queries, with respect to the documents in the collection.

Following this philosophy and depending on the particular task to be performed and language(s) to be used, the effectiveness of information retrieval systems participating in the CLEF campaigns is evaluated as follows:

- the collection containing the appropriate test documents is indexed and inserted into the system
- the sample queries are indexed and run using the system against the document index
- the results are evaluated based on the relevance assessments.

The TREC model has been adapted in CLEF to meet the specific requirements of a multilingual context. In particular, the initial experience gained from the cross-language track at TREC evidenced the need for native speakers of each language to be responsible for tasks such as topic creation and relevance assessment in that language. CLEF has thus adopted a distributed approach to the coordination of these tasks. This means that, for each new language added, it is necessary to find a group willing to be responsible for all the related language-dependent processing activities. The coordination of CLEF is thus a increasingly complex task, as it is necessary to ensure that each group follows carefully stipulated guidelines in order to guarantee the consistency and comparability of results over languages.

In the following sections, we describe the various tasks and test collections provided by CLEF and explain how the results of the participating systems are assessed and analysed

3 The Tasks

CLEF provides a series of evaluation tracks designed to test different aspects of information retrieval system development. The intention is to encourage systems to move from monolingual searching to the implementation of a full multilingual retrieval service. The design of these tracks has been modified over the years in order to meet the needs of the research community.

Previous to the launching of the 2002 campaign, we conducted a survey in order to acquire input with respect to the tasks to be offered. Two types of users were considered: cross-language technology developers and cross-language technology deployers. In particular, we solicited the opinion of previous CLEF participants. Their main recommendations can be summed up in the following list:

- Increase the size and the number of languages in the multilingual test collection (both with respect to documents and topics);
- Provide the possibility to test on different text types (e.g. structured data);
- Provide more task variety (question-answering, web-style queries, text categorization);
- Study ways to test retrieval with multimedia data;
- Provide standard resources to permit objective comparison of individual system components (e.g. groups using a common retrieval system can compare the effect of their individual translation mechanisms);
- Focus more on user satisfaction issues (e.g. query formulation, results presentation).

As far as possible, the findings of this survey were integrated into the definition of the CLEF 2002 campaign. Points that could not be taken up immediately will be considered for the future. Here below we describe the tracks and tasks offered by CLEF 2002.

3.1 Multilingual Information Retrieval

This is the main task in CLEF. It requires searching a multilingual collection of documents for relevant items, using a selected query language. Multilingual information retrieval is a complex task, testing the capability of a system to handle a number of different languages simultaneously and to merge the results, ordering them according to relevance. The multilingual collection for this track in CLEF 2002 contains English, German, French, Italian and Spanish documents. Using a selected topic (query) language, the goal is to retrieve documents for all languages in the collection, rather than just a given pair, listing the results in a merged, ranked list.

3.2 Bilingual Information Retrieval

In this track, any query language can be used to search just one of the CLEF target document collections. Many newcomers to CLIR system evaluation prefer to begin with the simpler bilingual track before moving on to tackle the more complex issues involved in truly multilingual retrieval. In the bilingual track of CLEF 2002, any topic language can be used to search target document collections in Dutch, Finnish, French, German, Italian, Spanish or Swedish. First-time CLEF participants only can choose to search the English document collection using a European topic language.

3.3 Monolingual (non-English) IR

Until recently, most IR system evaluation focused on English. However, many of the issues involved in IR are language dependent. CLEF provides the opportunity for monolingual system testing and tuning, and for building test suites in other European languages but not English. CLEF 2002 offers tasks for Dutch, Finnish, French, German, Italian, Spanish and Swedish.

3.4 Mono- and Cross-Language Information Retrieval for Scientific Texts

The rationale for this task is to study CLIR on other types of collections, serving a different kind of information need. The information which is provided by domain-specific scientific documents is far more targeted than news stories and contains much terminology. It is claimed that the users of this type of collection are typically interested in the completeness of results. This means that they are generally not satisfied with finding just some relevant documents in a collection that may contain much more. Developers of domain-specific cross-language retrieval systems need to be able to tune their systems to meet this requirement. See Gey & Kluck (2001) for a discussion of this point. In CLEF 2002, this track offers two distinct tasks:

- **AMARYLLIS:** System performance in searching a multi-disciplinary scientific database of French bibliographic documents will be studied. Tools are provided that can be used in the retrieval task (a controlled vocabulary in English and French
- **GIRT:** This task is based on the GIRT structured database of German social science documents. A German/English/Russian thesaurus and English translations of the document titles are available.

For each of the tasks listed above, the participating systems construct their queries (automatically or manually) from a common set of statements of information needs (known as topics) and search for relevant documents in the collections provided, listing the results in a ranked list.

3.5 Interactive CLIR

The aim of the tracks listed above is to measure system performance mainly in terms of how good the document rankings are. However, this is not the only issue that interests the user. User satisfaction with an IR system will be based on a number of factors, depending on the functionality of the particular system. For example, the way in which the results of a search are presented is of great importance in CLIR systems where it is common to have users retrieving documents in languages which they do not understand. When users are unfamiliar with the target language, they need a presentation of the results which will permit them to easily and accurately select documents of interest, discarding others. An interactive track that focused on this document selection problem was experimented with success in CLEF 2001 (see Oard & Gonzalo, forthcoming). In CLEF 2002 this track is being extended and will test both user-assisted query translation and as well as document selection.

4 The Test Collections

The main CLEF test collection is formed of sets of documents in different European languages but with common features (same genre and time period, comparable content); a single set of topics rendered in a number of languages; relevance judgments determining the set of relevant documents for each topic. A separate test collection is being created for systems tuned for domain-specific tasks.

4.1 Multilingual Corpus

The main document collection currently consists of well over 1,000,000 documents in eight languages – Dutch, English, Finnish, French, German, Italian, Spanish and Swedish. It contains both newswires and national newspapers for the period 1994-95. The CLEF 2000 collection contained documents in four languages: English, French, German and Italian. Spanish and Dutch were introduced for the first time in CLEF 2001 for different reasons. Spanish was included because of its status as the fourth most widely spoken language in the world. Dutch was added not only to meet the demands of the considerable number of Dutch participants in CLEF but also because it provides a challenge for those who want to test the adaptability of their systems to a new, less well-known language. Finnish has been included for CLEF 2002; its highly complex morphology and its membership of a different language family (Ugro-Finnic) with respect to the other European languages in the CLEF collection will provide an additional challenge for indexing. A Swedish collection of newswires is now in the process of being added.

We have established a set of criteria to determine whether a new language can be added to the multilingual collection. The language must be either of global importance (e.g. English, French, German, Spanish) or linguistically interesting (e.g. Finnish). Other important factors: the availability of the data and of a group willing to accept responsibility for language-dependent processing; the declared interest of a number of potential users.

Table 1 gives an idea of the contents and dimensions of the current multilingual collection. Not all the information is available yet for the 2002 additions.

Table 1: Sources and dimensions of the main CLEF document collection

Collection	Size (KB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens)	Median Size of Docs . (Features)
Dutch: Algemeen Dagblad	247141	106483	1282	166	112
Dutch: NRC Handelsblad	306207	84121	2153	354	203
English: LA Times	435112	113005	2204	421	246
French: Le Monde	161423	44013	1994	361	213
French: SDA	88005	43178	1683	227	137
German: Frank-furter Rundschau	327652	139715	1598	225	161
German: Der Spiegel	64429	13979	1324	213	160
German: SDA	147494	71677	1672	186	131
Italian: La Stampa	198112	58051	1915	435	268
Italian: SDA	87592	50527	1454	187	129
Spanish: EFE	523497	215738	2172	290	171
Finnish: Aamulehti	137000	55344	?	?	?
Swedish: TT - Tidningarnas Telegrambyrå	?	140000 (approx.)	?	?	?

SDA = Schweizerische Depeschenagentur (Swiss News Agency)
EFE = Agencia EFE S.A (Spanish News Agency)

Two distinct scientific collections are also available: the GIRT database of about 80,000 German social science documents, which has controlled vocabularies for English-German and German-Russian, and – new for CLEF 2002 - the Amaryllis multidisciplinary database of approximately 150,000 French bibliographic documents and a controlled vocabulary in English and French.

4.2 Topics

The participating groups derive their queries in their preferred language from a set of topics created to simulate user information needs. Following the TREC philosophy, each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. The English version of a typical topic from CLEF 2001 is shown below:
Title: U.N./US Invasion Haiti
Description: Find documents on the invasion of Haiti by U.N./US soldiers.
Narrative: Documents report both on the discussion about the decision of the U.N. to send US troops into Haiti and on the invasion itself. They also discuss the direct consequences.

The title contains the main keywords, the description is a “natural language” expression of the concept conveyed by the keywords, and the narrative adds additional syntax and semantics, stipulating the conditions for relevance assessment. The motivation behind these structured topics is to provide query “input” for all kinds of IR systems, ranging from simple keyword-based procedures to more sophisticated systems supporting morphological analyses, parsing, query expansion and so on. In the cross-language context, the transfer component must also be

considered, whether dictionary or corpus-based, a fully-fledged MT system or other. Different query structures may be more appropriate for testing one or the other methodology.

For CLEF 2002, 50 such topics have been developed on the basis of the contents of the multilingual collection and topic sets have been produced in all eight document languages. Additional topic sets in Portuguese Russian, Japanese, Chinese have been prepared, other languages may be offered depending on demand. The same topic set is used for the multilingual, bilingual and monolingual tasks. Participants can thus choose to formulate their queries in any one of at least ten European or two Asian languages. Separate topic sets are developed for the scientific collections: in German, English and Russian for the GIRT task, and French and English for Amaryllis.

4.3 Relevance Judgments

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead, approximate recall figures are calculated by using pooling techniques. The results submitted by the participating groups are used to form a "pool" of documents for each topic and for each language by collecting the highly ranked documents from all the submissions. The assumption is that if a sufficient number of diverse systems contribute results to a pool, it is likely that a large percentage of all relevant documents will be included. All documents not included in the pool remain unjudged and are therefore assumed to be irrelevant. A main concern with such a pooling strategy is that if the number of not detected relevant documents is above a certain (low) threshold, the resulting test collection will be of limited future use in testing systems that did not contribute to the pool. A grossly incomplete pool would unfairly penalize such systems when calculating precision and recall measures. This pooling strategy was first adopted by TREC and has been subsequently employed by both NTCIR and CLEF. A number of studies have been made to test its validity (see Zobel, 1998; Voorhees, 2000).

A test of the completeness of the pools used for the CLEF 2000 campaign can be found in Braschler (2001). The test reveals that the completeness of the relevance assessments compares favorably to that of the assessments used for previous TREC ad-hoc campaigns. As already mentioned, relevance assessment of the documents in the pool is distributed over a number of different sites and performed in all cases by native speakers. The results are then analyzed centrally using recall and precision measures and run statistics are produced and distributed.

The problems involved in multilingual topic creation and relevance assessment are discussed in more detail in Kluck & Womser-Hacker (2002).

5 Results Analysis

The CLEF campaign evaluates all official submissions based on the relevance assessments. A variety of measures are calculated both for every individual submission and for overall statistics. The two central evaluation measures used are Recall and Precision. Recall measures the ability of a system to present all relevant items, whereas Precision measures the ability of the system to present only relevant items.

$$\text{Recall } \rho_r(q) := \frac{\left|D_r^{rel}(q)\right|}{\left|D^{rel}(q)\right|}$$

and

$$\text{Precision } \pi_r(q) := \frac{\left|D_r^{rel}(q)\right|}{\left|D_r(q)\right|},$$

where $D_r(q) := \{d_1,...,d_r\}$ is the answer set to query *q* containing the first *r* documents. The choice of *r* depends on the preference of the user: a low value for *r* implies that the user is interested in few, high-precision documents, whereas a high value for *r* means that the user conducts an exhaustive search. $D^{rel}(q)$ is the set of all relevant documents, and $D_r^{rel}(q) := D^{rel}(q) \cap D_r(q)$ is the set of relevant documents contained in the answer set (Schäuble 1997).

The two measures are somewhat in conflict: it is desirable in most cases to optimize for both measures, i.e. retrieving a maximum of relevant items while retrieving a minimum of irrelevant items, but systems that optimize for better recall often do so at the expense of precision, while systems that optimize for precision often adopt a conservative retrieval strategy that leads to lower recall. It is therefore important to analyze system performance in a variety of scenarios, such as precision at low recall levels, recall at low precision levels and balance between precision and recall. In CLEF, precision figures for a range of recall levels are published, as well as the popular average precision measure, which summarizes performance across various recall levels.

Graphically, precision figures at multiple levels of recall can be visualized in the form of a recall precision graph (see Figure 1).

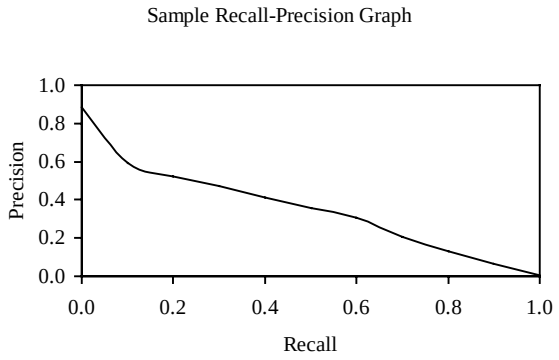


Figure 1: Sample CLEF Recall-Precision Graph

In the case of one particularly popular application for retrieval technology, the World Wide Web, recall is often seen as of secondary importance, since for most search requests there is an overwhelming number of potentially relevant hits. In such scenarios, high precision at low recall levels is increasingly popular as an evaluation metric. The CLEF policy of publishing a range of performance measures also caters for this application.

The goal of CLEF is a comparative evaluation of retrieval techniques. Absolute performance levels do not generally carry over across different experimental setups. CLEF facilitates result comparison by publishing both graphs that summarize overall results and by presenting a graphical comparison of individual results to median performance (see Figure 2).

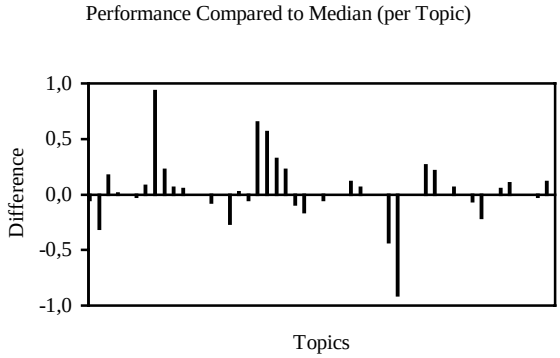


Figure 2: Comparison to Median by Topic

Starting with the 2001 campaign, CLEF also publishes an analysis of the statistical significance of performance differences observed between submissions by different participants (Braschler, forthcoming). Preliminary figures suggest that it is hard to achieve statistically significant performance differences, since the variability in performance between queries tends to be higher than the variability of performance between systems. Similar observations have been made before for TREC experiments (Tague-Sutcliffe & Blustein, 1995). CLEF tries to address this problem by producing topic sets that can be combined with earlier years' campaigns into larger sets, which helps to obtain more reliable figures for post-campaign evaluations. Additionally, CLEF also publishes average precision figures for individual experiments per topic, allowing comparison of systems for specific topics.

6 Test-suites

A final product of an evaluation campaign, or a series of campaigns, is a set of reusable test collections. This consists of the data collections, topic sets and relevance assessments for each task. These test-suites can be used - together with the trec_eval package (ftp://ftp.cs.cornell.edu/pub/smart/) available from Cornell University - by developers to test and tune their systems, independently of the evaluation campaigns. Unfortunately, for copyright reasons, this valuable resource is currently only available to registered participants. An objective of the CLEF project is to stipulate agreements with the data providers that will make the test-suites produced by the evaluation campaigns also accessible to the wider R&D community for benchmarking purposes.

7 Summing Up

An evaluation activity of this type is important in that provides a forum in which traditional state-of-the-art methods can be compared against new techniques. The first two CLEF campaigns proved very successful. Over thirty groups from both academia and industry participated in CLEF 2001, up more than 50% with respect to the previous year. CLEF 2002 is now under way.

Participating groups in CLEF 2001 attempted both traditional and innovative approaches to CLIR. All kinds of source to target transfer mechanisms were employed, including both query and document translation. Commercial and in-house resources were used and included machine translation, dictionary and corpus-based methods. The search strategies used varied from traditional IR to a considerable employment of natural language processing techniques. Different kinds of query expansion techniques were tested. Different groups focused on various aspects of the overall problem, ranging from the development of language-independent tools such as stemmers to much work on language-specific features like morphology and compounding. A number of groups compared different techniques in different runs in order to evaluate the effect of a given technique on performance. In particular, it was noticeable that many groups were testing systems that integrated more than one translation method, e.g. MT or bilingual dictionary look-up combined with a data extracted from a comparable of parallel corpora. A complete record of the CLEF 2001 experiments and a detailed analysis of the results can be found (Peters et al, forthcoming). A preliminary description is already available in the Working Notes which can be accessed on our Web site (http://www.clef-campaign.org/).

In future years, we hope to go further in the extension of CLEF evaluation tasks, moving gradually from a focus on cross-language text retrieval and the measuring of document rankings to the provision of a comprehensive set of tasks covering all major aspects of multilingual, multimedia system performance with particular attention to the needs of the end-user.

Acknowledgments

We gratefully acknowledge the support of all the data providers and copyright holders:
The Los Angeles Times, for the English data collection;
Le Monde S.A. and ELDA: European Language Resources Distribution Agency, for the French data;
Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections;
InformationsZentrum Sozialwissenschaften, Bonn, for the GIRT database;
Institute de Information Scientifique et Technique, Vandoeuvre, France, for the Amaryllis database;
Hypersystems Srl, Torino La Stampa, for the Italian data;
Agencia EFE S.A. for the Spanish newswire;
NRC Handelsblad; Algemeen Dagblad - PCM Landelijke dagbladen/Het Parool for the Dutch collections;
Aamulehti Oyj for the Finnish newspaper data;
Tidningarnas Telegrambyrå (TT) SE-105 12 Stockholm, Sweden for the Swedish data;
Schweizerische Depeschenagentur (SDA) for French, Italian and German newswire data.
Without their assistance, the CLEF evaluation activity would be impossible.

References

Braschler, M. (forthcoming). CLEF 2001 – Overview of Results. In In Peters et al. op cit.

Cleverdon, C. (1997). The Cranfield Tests on Index Language Devices. In K. Sparck Jones and P. Willett (Eds.). Readings in Information Retrieval, pp 47-59. Morgan Kaufmann, 1997.

Gey, F.C. & Kluck, M. (2001). The Domain-Specific Task of CLEF – Specific Evaluation Strategies in Cross-Language Information Retrieval. In C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science 2069, Springer Verlag, pp 48-56.

Gey, F.C. & Oard, D.W. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries. NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001).

Kando, N., Aihara, K., Eguchi, K., Kato, H. (Eds.) (2001). Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, National Institute of Informatics (NII), ISBN 4-924600-89-X.

Kluck, M. & Womser-Hacker, C. (forthcoming). Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. Paper to be published in LREC2002 Proceedings.

Oard, D.W. & Gonzalo, J. (forthcoming). The CLEF 2001 Interactive Track. In Peters et al. op cit.

Peters,C., Braschler, M, Gonzalo, J., Kluck, M. (Eds.). Proceedings of CLEF 2001. Lecture Notes in Computer Science, Springer Verlag, in print.

Schäuble, P.(1997). Content-Based Information Retrieval from Large Text and Audio Databases. Section 1.6 Evaluation Issues, Kluwer Academic Publishers, pp 22-29.

Tague-Sutcliffe, J. & Blustein, J. (1995). A Statistical Analysis of the TREC-3 Data. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226. p385ff.

Voorhees, E.M. (2000). Variations in relevance judgments and the measure of retrieval effectiveness. Information Processing and Management (36), pp 697-716.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, pp 307-314.

Evaluating Efficiency vs. Effectiveness for Vague Queries and Similarity Search in Digital Libraries

Norbert Fuhr*
University of Dortmund, Germany

Abstract

In this paper, we discuss the relationship between result quality and effectiveness, thus motivating approximation methods for computation-intensive searches. For evaluating the quality of approximations, we first propose a new paradigm for measuring the quality of content-based searches. For standard evaluations, this new paradigm yields metrics similar to known measures. However, if the user is only interested in a small number of documents, we get different metrics. Based on our paradigm, we derive a new measure for evaluating the quality of approximation methods for similarity searches. Finally, we discuss the relationship between similarity-based and relevance-based evaluations, and we justify the need for the latter.

1 Introduction

Most searches in digital libraries (DL) are content-oriented. Thus, a DL system has to cope with the intrinsic uncertainty and vagueness of this type of searches. Since a wide range of models and methods can be applied for dealing with these issues, proper evaluation methods are required in order to draw conclusions about the effectiveness of a specific method or system.

Content-oriented searches have been studied for a long time in the field of information retrieval. However, most research in this area has focused on text documents, for which both efficient and effective methods have been developed. With the increased availability of multimedia documents, DL systems also should provide search tools for non-textual content. Looking at the state of the art of multimedia retrieval, one can see that most methods developed here are feature-based. Since most approaches focus on similarity searches (i.e. find media objects similar so a given one), we will mainly consider this type of searches in the remainder of this paper[1]. Unfortunately, the feature spaces for media objects are either high-dimensional or only metric spaces, thus similarity searches in these spaces are rather time-consuming, which causes high response times.

In this situation, an evaluation focusing on the quality of the final result may not be appropriate. As a more suitable approach, we should regard effectiveness, which also takes into account the effort of the user for reaching the present result. Roughly speaking, effectiveness can be interpreted as the following ration:

$$\text{effectiveness} = \frac{\text{result quality}}{\text{user's effort}}$$

Thus, a high-quality result is less effective in case the user has to wait very long for the system's answer. Given the fact that many multimedia searches are highly interactive, response time is a crucial factor.

In order to overcome the problem of time-consuming searches in feature spaces, several researchers have proposed methods for approximate similarity searches ([Zezula et al. 98], [Weber et al. 98], [Arya et al. 98]). In this case, the system does not compute the precise result according

*Email: norbert.fuhr@acm.org
[1]However, most claims of this paper also hold for vague queries in general.

to the definition of the similarity metrics. Instead, the system focuses on that part of the result that can be computed most efficiently, thus missing some of the most similar objects. This approach raises the problem of proper evaluation: How can we measure the loss of retrieval quality that is caused by the approximation? Most researchers propose some metrics for measuring the deviation from the precise result. However, since a user's perception of similarity may be very different from the metrics implemented in the system, one should also consider an information retrieval oriented approach based on the notion of relevance.

In this paper, we try to link these two evaluation approaches. First, we propose a new paradigm for measuring the quality of content-based searches, and we discuss a number of possible applications. Then we use this paradigm for the defining a new evaluation measure for approximate k-nearest-neighbor searches. Finally, we discuss the relationship between similarity-based and relevance-based metrics.

2 A new quality paradigm

As motivating example for defining a quality metrics, consider the following situation: We have a group of users who want to look at a single document for each query[2]. There are two groups of queries (I and II), with different numbers of relevant documents in the database: for I, there are 10 relevant documents per query, whereas for II, the collection contains 100 relevant documents in each case. Now we have two systems A and B, where we process query group I with system A and the queries from group II with system B. Each of the systems succeeds in exactly 50% of all cases (i.e. the only document retrieved is a relevant one).

Now we would like to know which of the two systems is better: Since they both succeeded in 50 % of the cases, one could conclude that they are of equal quality. On the other hand, it is obvious that the tasks have different complexity: finding one out of 10 relevant documents is much harder than one out of 100. So, in order to make a fair comparison, we need a measure for the complexity of the task.

For this purpose, we propose the following approach: We assume that there is an ideal system which has complete knowledge about the query and the documents. Thus, by comparing document and query, the system is able to decide whether or not the document is relevant. Furthermore, let us assume that there is no access structure, so the ideal system has to do a linear scan through the database in order to find relevant documents. Based on this ideal system, we define task complexity as the expected number number of processing steps for performing the requested task in a perfect way.

As a simple example, assume that there is only one relevant document in the database; in the average case the ideal system will have to consider half of the documents in order to identify the relevant one.

For our first example from above, computation of task complexity is a bit more difficult. Assume that we have a collection consisting of r relevant and s non-relevant documents. Now the ideal system does a linear scan until j relevant documents have been found. For the corresponding expected number of non-relevant documents that have to be considered, [Cooper 68] (see also [Raghavan et al. 89]) has derived the measure of "expected search length" as follows:

$$\mathrm{esl}(j, r, s) = \frac{j}{r+1} s \tag{1}$$

In the following, let us assume that $N = r + s$ denotes the collection size and that $r \ll N$, so $N \approx s$. Thus we would have a task complexity of $\frac{1}{11}N$ for the query group I and of $\frac{1}{101}N$ for the second group, so system A performs a much harder task.

We could regard the task complexity defined above as a measure of absolute quality. For most evaluations, however, such a measure is not required. Typically, evaluations use connected samples, e.g. the same set of queries is processed by all systems to be compared.

[2]This is a simplification of today's Web search, where typical users regard the top 10 documents only.

So most evaluations need a measure of relative quality only. For example, assume that we want to evaluate Web search engines. Since most users regard the first ten answers only, a realistic evaluation should be restricted to these documents. Assuming that there are $r \geq 10$ relevant documents on the Web, the ideal system would retrieve 10 of them, thus yielding an absolute quality of $\frac{10}{r+1}N$. For a Web search engine retrieving only k relevant documents, the task complexity would be $\frac{k}{r+1}N$; thus the relative quality of the latter system is $k/10$. (In this case, our measure for relative quality corresponds to the well-known precision measure.) So we can define our measure of relative quality as follows:

$$\text{relative quality} = \frac{\text{task complexity of solution of real system}}{\text{task complexity of perfect solution}}$$

Besides stopping after a certain number of documents retrieved, another typical user standpoint depends on the number of relevant documents retrieved. This standpoint is reflected in the well-known recall-precision graphs, where precision at certain recall points (the ratio of all relevant documents retrieved) is plotted. Here we only consider the absolute number of relevant documents retrieved, which we denote by k. Whereas the ideal system would retrieve only those k relevant documents (assuming $k \leq r$), the real system will retrieve n documents until k relevant ones have been located. In order to compute the relative quality in this case, we use the following idea: The real system performs a task that is less complex than that of the ideal one: Whereas the ideal system has to retrieve k relevant ones from a total of r, the real system misinterprets some non-relevant documents as relevant ones; assuming the same ratio of misinterpretation for the whole collection as for the retrieved documents, we can conclude that the real system regards $r \cdot n/k$ documents in the collection as being relevant. Thus, the relative quality of the real system is

$$\frac{\text{esl}(k, r \cdot n/k, N)}{\text{esl}(k, r, N)} = \frac{\frac{k}{r \cdot n/k+1}N}{\frac{k}{r+1}N} = \frac{r+1}{r \cdot n/k + 1} \approx \frac{k}{n}$$

Obviously, for small values of r, the last approximation is not appropriate; in this case our measure of relative quality deviates from the standard precision measure that would be applied usually.

The last type of measurement can also be applied for evaluating approximate k-nearest-neighbor (kNN) searches. From the k documents returned by approximate search, we consider the highest rank (which should be k in the ideal case) for evaluation[3]. Let n denote the highest rank, then our argument goes as follows: Instead of retrieving exactly the top k documents from the collection, the approximation method selects k documents from the top n ones. Thus, the relative quality of the approximation method is

$$\frac{\text{esl}(k, n, N)}{\text{esl}(k, k, N)} = \frac{\frac{k}{n+1}N}{\frac{k}{k+1}N} = \frac{k+1}{n+1}$$

This formula tells us that we should consider the *relative* difference in the rank numbers (+1) as quality measure. For example, if the user is looking for the single most similar object, the second best answer would represent a quality of 2/3, whereas an offset of 1 for the 10 most similar objects would yield a quality of 11/12.

3 Similarity-based vs. relevance-based evaluation

Evaluation of approximate kNN search as described above is based on the similarity metrics implemented in the system. However, this type of evaluation tells us nothing about a user's impression of the quality of the result. A user's general perception of similarity may be quite different from the system's similarity metrics. Furthermore, in many cases, similarity may be a

[3]It also would be possible to consider apply the method described here to all ranks (and then compute the average), but we restrict to the simple case in this paper.

very poor tool (but the best one provided by the system) for supporting the current information need (e.g. searching for images containing certain objects). Thus, quality judgments by the user must be considered in order to achieve meaningful evaluation results.

As starting point, we choose the standard information retrieval approach of a binary relevance scale. So a user judges about the relevance of single answer objects with respect to her/his information need.

For illustrating the relationship between similarity-based and relevance-oriented evaluation, the rank-recall graphs in figure 1 show the results of two queries from the TREC collection[4]. These graphs depict the relationship between the number of documents retrieved and the number of relevant documents among them. As user standpoint, we assume that (s)he is looking for a certain number of documents retrieved; thus, we choose precision as appropriate retrieval measure. Graphically, precision can be interpreted as follows: for any point on the curve, the corresponding precision is the slope of a straight line from the origin to this point. Applying this interpretation on the two graphs, it can be seen that the result illustrated on the left in figure 1 has constant precision for the first 13 documents, after which precision goes down rapidly. In contrast, for the result shown in the right half of figure 1, precision starts decreasing very early, but goes down rather smoothly.

Now let us assume that we have a system performing approximate searches. Such a system would miss some of the top-ranking documents. Using a similarity-based evaluation, any such miss would be interpreted as a loss of quality. For relevance-based evaluation, however, a potential loss of quality due to approximation depends on the shape of the rank-recall curve: For the graph shown in the left half of figure 1, assume that the system operates in the range of constant slope; in this case, missing some of the documents would have no effect on retrieval quality. Only when we get into the range of decreasing slope — because the system misses to many documents, or the user want to see more documents — retrieval quality is affected by approximation. In contrast, the rank-recall curve depicted on the right tells us that any approximation will also result in a loss of retrieval quality.

From these examples, it can be seen that the findings from similarity-based evaluation can not be extrapolated to relevance-based evaluation. So the former tells us very little about the quality of a system from a user's point of view. The good news is that approximate searching does not necessarily reduce retrieval quality. Obviously, relevance-based evaluation is required to judge about the quality of an approximation method.

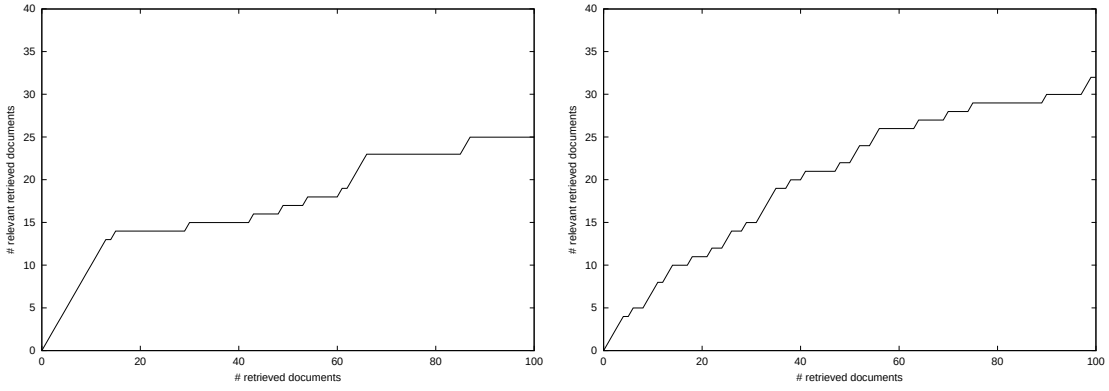


Figure 1: Rank-recall graphs of two example queries

[4] `http://www.trec.nist.gov`

4 Conclusions and outlook

In this paper, we have proposed a new paradigm for measuring the quality of content-based searches, and we have discussed its application to different situations. The final comparison of similarity-based and relevance-based evaluation has justified the need for the latter.

The evaluation paradigm proposed has been justified from a system-oriented view. Further research should investigate its relationship to the user oriented view: Do users also estimate the value of an item as being proportional to its relative frequency, or are other scales more appropriate (e.g. information theory would suggest a logarithmic scale)?

Here we have discussed only fairly simple evaluation problems, regarding a database as a set of media objects from which the interesting ones are to be selected. In the future, we will investigate the application of this paradigm to more complex situation, e.g. the retrieval of relevant parts of structured documents.

In digital libraries, we typically have a rich structure, with various classes of objects and links of different types in between. In this situation, users may reach the interesting objects via different paths. A user-oriented evaluation may be based on simple measurements such as the time needed for specific task or success rate. For system-oriented researchers and developers, however, the quality of certain components is more relevant, in order to make develop better methods or make good design choices. Thus, there is still the need for proper evaluation methods in this area.

References

Arya, S.; Mount, D.; Netanyahu, N.; Silverman, R.; Wu, A. (1998). An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM 45(6)*, pages 891–923.

Cooper, W. (1968). Expected Search Length: A Single Measure of Retrieval Effectiveness Based on Weak Ordering Action of Retrieval Systems. *Journal of the American Society for Information Science 19*, pages 30–41.

Raghavan, V.; Bollmann, P.; Jung, G. (1989). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information Systems 7(3)*, pages 205–229.

Weber, R.; Schek, H.-J.; Blott, S. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In: *Proceedings of the 24th International Conference on Very Large Databases*, pages 194–205. Morgan Kaufmann, Los Altos, California.

Zezula, P.; Savino, P.; Amato, G.; Rabitti, F. (1998). Approximate Similarity Retrieval with M-Trees. *VLDB Journal 7(4)*, pages 275–293.

Digital Library Attributes: Framing Research and Results

Robert J. Sandusky
(sandusky@uiuc.edu)
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St.
Champaign, Illinois 61820

Introduction

This paper introduces a framework for evaluating digital libraries of various kinds. The framework consists of six attribute groupings (audience; institution; access; content; services; design and development). Each attribute grouping contains a number attributes expressed as continua, or dimensions. This framework allows us to think about digital libraries in a flexible and holistic manner and helps us compare digital libraries attribute by attribute. Digital libraries are proliferating and are being built and operated by a wide variety of institutions for a wide variety of reasons. The motivation for the development of this framework is to help us understand and apply what has been learned from research on digital libraries to current and future digital library evaluation or development efforts. By understanding and clearly expressing how two digital libraries are similar or different, we should be able to better apply lessons learned from previous evaluation or user study efforts. In addition, this analysis can highlight aspects of digital libraries that have been under- or un-examined in previous studies.

The framework may also prove valuable in helping focus digital library evaluation efforts by providing us with language to articulate the purpose and context of our work. For example, an evaluation project can be organized to examine the effectiveness of a particular digital library in terms of one or more of the attributes. Or, a digital library could be evaluated in terms of how variation in along one dimension (e.g., governance) affects another (e.g., content coherence). These are examples of evaluating a single digital library (*within-DL* evaluation). The framework can play a similar role in comparisons of multiple digital libraries. For example, a researcher might want to examine how different approaches to content sourcing affect content fit (*between-DL* evaluation).

The framework presented here is provisional and is based upon review of existing digital library literature. The utility of the framework as a means of organizing research is being tested in an evaluation of a commercial digital library.

What kinds of digital libraries exist?

There are many DLs available on the Internet and on private networks or intranets. I assert that while there are many DLs, most of the DLs that have undergone careful evaluation have not been representative of the full range of existing DLs. If this assertion is accurate, how can we begin to understand what distinguishes DLs?

Borgman makes a distinction between ‘research’ and ‘practice’ perspectives regarding DLs. Her examination of competing definitions of the term ‘digital library’ focuses on accommodating both perspectives (Borgman, 2000). Lesk examines a handful of early projects in detail, all of which are academic in nature and motivated by the interests of researchers (Lesk, 1997). These two perspectives do not acknowledge other kinds of DLs that exist and continue to be developed.

Lynch suggests that there are three types of digital libraries. Commercial systems targeted for specific professions or disciplines, like Westlaw, Lexis or Nexis; research systems (he cites, for example, the systems funded by the NSF/ARPA/NASA Phase I Digital Libraries Initiative as examples (NSF)); and systems embedded within existing research and academic libraries. He emphasizes distinguishing 'active' from 'passive' DLs. Passive DLs support locating and reviewing information and active DLs also include elements to support active work (analysis and collaboration) (Lynch, in press).

Lynch also mentions retail systems like Amazon.com and, more generally, consumer health information systems and investment and stock trading systems as examples of active commercial DLs. They are important, he argues, precisely because they are *active* DLs that support decision making by their users and generally modify in some way individual and / or social behavior through their inclusion of analytic and collaborative tools. He also notes important differences between commercial DLs and DLs arising out of the traditions of librarianship. Commercial, active DLs lack a commitment to preservation and continuity of access to content. They are also sensitive to marketplace dynamics because a commercial DL's audience may turn quickly to a competitor's DL if the competitor provides advantages in usability, content or cost.

Attributes of Digital Libraries

Rather than try to identify the three, four, five or whatever number of 'general kinds' of DLs that exist, it's more useful to think about the general attributes of any DL in order that we have a flexible vocabulary to use when examining DLs. Some researchers have already begun to address this issue. Saracevic presents a holistic, six-level framework for classifying and thinking about evaluations of information retrieval systems (Saracevic, 1995). The levels are (1) engineering, (2) input, (3) processing, (4) output, (5) use and user and (6) social. He later adapts this framework for application to digital libraries (Saracevic & Covi, 2000), presenting a modified framework consisting of these seven levels: (1) social, (2) institutional, (3) individual, (4) interface, (5) engineering, (6) processing and (7) content. The seven levels are grouped as 'user-centered' (levels 1-3) and 'system-centered' (levels 5-7), with level 4 as a boundary area between the user and the system. In either scheme, each level represent an aspect of the system that can be subjected to evaluation.

Another holistic description scheme, consisting of four major dimensions is presented by Fuhr et. al. Their dimensions are (1) data/collection, (2) system/technology, (3) users and (4) usage (Fuhr, Hansen, Mabe, Micsik, & Sølvberg, 2001). Their scheme is an attempt to define a comprehensive set of evaluation criteria along with some suggestions for the metrics to be used in an evaluation.

The framework proposed here is different from those presented in the work described above in three ways. First, this framework seeks to make the technical, the social and the organizational attributes of digital libraries visible without giving any of these areas primacy. Second, the conceptualizations of the attributes contained here are purposely inclusive. For example, in contrast to Fuhr et. al., access here includes issues influenced by DL policies, such as the persistence of the content contained in the DL. This is consistent with how others have conceptualized access. Borgman, for example, defines access broadly: "…as connectivity to a computer network and to available content, such that the technology is usable, the user has the requisite skills and knowledge, and the content itself is in a usable and useful form (Borgman, 2000)." Rice et.al., in a recent book, present a multi-dimensional framework of access that takes physical, cognitive, affective, economic, social, political and mediative aspects into account (Rice, McCreadie, & Chang, 2001). Finally, this framework, in contrast to the emphasis on layers presented by Saracevic, leaves open the possibility of considering how varying certain attributes may affect other DL attributes. This is similar to the 'cascades of interactions' model proposed by Bates (Bates, 2002).

I have developed a list of attributes, expressed as continua or dimensions, to help identify similarities, differences and patterns among different DLs. This system of attributes contains six attribute groupings:

- *Audience* is concerned with attributes of the targeted and actual users of a particular DL
- *Institution* refers to the library, university, company or other entity that sponsors the creation of a DL
- *Access* refers to who can use the DL, under what conditions; also the features the DL provides to support access to content
- *Content* refers to the information (e.g., documents) contained in the DL
- *Services* is concerned with human or technology-based capabilities that link 'collections to those using them and link people to one another' (Bishop & Star, 1996)
- *Design and development* is concerned with the process of building and maintaining a DL

The groupings and their associated dimensions are summarized in Table 1. The labels anchoring each dimension are shown in **bold** text.

Digital Library Attributes		
Audience	attributes of the targeted and actual users of a particular DL	
	Scope	Is the target user community **constrained** or **unconstrained**? For example, only the employees of one company can use an internal DL in a constrained setting, or only the students, faculty and staff of one university can use certain DL services provided by a campus library. Other DLs may require fees and others may be free and available on the open Internet (subsidized by advertising or government funding).
	Coherence	Does the main user community consist of people who have a set of similar interests or is the user community the general public? A **coherent** audience consists of intended or actual users who have key attributes in common relative to the DL. A DL designed to support researchers in astrophysics has a coherent audience. A **diffuse** audience has few attributes in common (or only by chance). A DL intended for the general public has a diffuse target audience.
	Fit	Are the actual users of the DL the same as the target users? A DL with **mis-fit** has actual users who are not members of the target audience. A DL with **close fit** has users who are members of the target audience.
Institution	the library, university, company or other entity that sponsors the creation of a DL	
	Governance and control	Who and what drives decision-making with regard to the DL's design, its operations and its ongoing development (Lynch, in press)? A competitive, commercial DL has **closed** governance and control in order to minimize leakage of plans and proprietary information to competitors. A government supported DL typically has **open** governance and control.
	Type	Is the sponsoring institution **public** (government, university) or **private** (a corporation)?
	Economic model	How is the DL funded and sustained? A DL supported only by **grants** supplied by a foundation or similar agency would be placed on one end of the dimension. A DL supported only by the direct (pay for use) or indirect (advertising) generation of revenue based on use would be placed at the **sales** end of the dimension.
	Mission	What is the stated purpose of the DL? An **experimental** DL is developed primarily to examine the feasibility of new techniques. A **production** DL is developed to provide a needed information service to a community of users.
Access	who can use the DL, under what conditions; also the features the DL provides to support access to content	
	Payment Model	Do DL users pay directly or indirectly to us the DL? **Fee** based DLs might implement 'pay per view' or monthly subscriptions. In either

		case, the correlation between payment and access is direct and obvious. Less direct payment models include access to a DL as a side effect of paying tuition; a federal DL supported by tax monies; a commercial DL supported by advertising revenue. These DLs would belong near the **free** end of the dimension. In these cases, the connection between payment and use is less obvious. The user incurs no more cost if he or she decides to use the DL than if he or she does not use it.
	Visibility	Is the DL itself visible on the Internet (whether or not any or all of the materials are available for free) or is the DL's very existence hidden from view (by firewall or other security technology)? Private DLs hosted on carefully controlled intranets are **hidden** unless specific privileges are granted. Commercial and governments DLs accessible through the public network are **visible**.
	Persistence	Are the DLs contents guaranteed or implied to be available in the future? A DL containing academic papers would be expected by its audience to be **preserved** forever in some format (for decades, if not centuries). The context of DL development and use implies preservation of content and continuity of access. The contents of a commercial DL are unlikely to be preserved by its provider: continuity of access is not implied. Their content is **ephemeral**.
	Coherence	Is access to the DL content enhanced through the application of principles of information organization? (Given the previously stated definitions of a DL, some organization must be available.) A minimally organized (or **unorganized**) DL might only provide a full-text search or a single-level topic navigation structure. An **organized** DL might provide multiple types of search access and multiple complex topic structures. See also the attribute 'organization' within the grouping 'Content'.
Content	the information (e.g., documents) contained within the DL	
	Scope	How extensive is the topical coverage of the content included in the DL? A DL with **limited** content scope would have only information about a single topic. A **comprehensive** DL would contain information pertaining to any branch of human knowledge.
	Fit (to audience needs)	Does the content suit the intended or actual audience? A science DL intended for grade school children would emphasize introductory, easily read information and would have **close fit** with the audience. The same collection would not be a good fit for scientists conducting basic research (**mis-fit**).
	Coherence	Is the content of the DL collected in a way to provide coherent coverage of any included topic? A **coherent** DL might provide a range of introductory and specialized information in a variety of media, treatments and languages. A **non-coherent** DL would provide some, but not all, parts of these ranges. Significant gaps might be found in a non-coherent DL.
	Organization	Is the DL content organized in a way that facilitates information finding and access? A minimally organized (or **unorganized**) DL might only provide a full-text search or a single-level topic navigation structure. An **organized** DL might provide multiple types of search access and multiple complex topic structures. See also the attribute 'coherence' within the grouping 'Access'.
	Specialization	Is the DL content **specialized** for a specific audience, like astrophysicists, or is the content **general** or wide ranging in nature?
	Digitized content	Is the content in the DL digitized from existing, non-digital material or is the content created as digital information? A DL of current scientific information would likely contain **original digital** content. A DL of nineteenth century newspapers would contain **digitized** versions of content originally produced in another medium.
	Sourcing	Is the DL content from a single or multiple sources? **Single-sourced** DLs would present content created by a single entity or a single, controlled process (e.g., a journal or periodical with a defined editorial process). **Multiple-sourced** DLs would present

		content created by many entities.
	Restricted	Is the content in the DL in the **public domain** (not subject to copyright restrictions) or **restricted** by copyright or similar restrictions on intellectual content (implying that rights management is an important concern)?
Services	human or technology-based capabilities that link 'collections to those using them and link people to one another' (Bishop & Star, 1996).	
	Nature of Interaction	This dimension represents the manner in which the DL does or does not provide support for active work (Lynch, in press). Those DLs that support simple access to information (similar to the basic model of a physical library) are **passive**. DLs that provide analytical tools or user-to-user (or user-to-expert) collaboration are **active**. See specific types of interaction 'Analytic', 'Collaborative' and 'Reference', below.
	Analytic	Does the DL provide support for evaluation or assimilation of information? **Analytic** DLs include interactive tools to support data analysis or ability to store user-generated data in the context of the DL. **Non-analytic** DLs do not include these facilities.
	Collaborative	Does the DL provide support for communication between DL users? **Collaborative** DLs often include embedded chat or message board systems; defined interest groups; user-to-user awareness facilities; etc. **Non-collaborative** DLs would not include these facilities.
	Collection	Does the DL provide content selection and organization services? A DL with clear collection development and indexing policies would be **controlled**. A DL allowing the addition of content without an evaluative process would be **uncontrolled**. A simple example is to contrast moderated with un-moderated mailing lists or message boards.
	Reference	Does the DL give users the ability to interact with domain and / or DL experts? Can users contact someone via email? In real-time (e.g., technical support via chat interface)? This is a specific type of collaboration service. DLs either **include** or **lack** this service.
Design and Development	the process of building and maintaining a DL	
	Design approach	Is the DL designed with the issues of audience, institution, access and content in mind? An **experimental** approach is often employed with systems driven by technical ideas: innovations in database, network, interface or software design or ideas interesting to the DL designers and developers. **Socially grounded** approaches take some or all of the issues of audience, institution, access and content into account from the outset. Commercial DLs intended to succeed in a competitive environment take at least some social issues into account.
	Design cycle	A **non-iterative** DL would be one that was developed and left relatively unchanged after one design iteration (e.g., a DL built with funding by a single grant or a DL that fails to generate a means of sustaining itself). An iteratively designed DL undergoes multiple, regular revisions based upon a methodical approach that includes feedback from the analysis of DL use (**iterative**).
	Consulted users	Who stands in for / speaks for the intended DL users during DL design and development? Many DLs are conceived and developed without direct input from users or their proxies despite the higher risk of mis-fit (**none**). In other cases, some kind of proxy user may be consulted by the designers and developers because the actual users are not easily contacted. The other extreme is represented by a project where **actual users** are included as full members of the design and development group (e.g., participatory design).

Table 1 - Digital Library Attributes

The attributes in Table 1 can be used to identify the kinds of DLs that exist and the kinds of DLs that have and have not been subject to evaluation. There are, however, limitations to how these attributes can be used to analyze DLs. It may be impossible to learn about the details of the collection development policies or the kind of systems development methods employed in the construction and maintenance of a particular DL. The placement of any particular DL along any of these continua is also subject to debate. These continua may be helpful in comparing a small number of DLs on a small number of dimensions. So it might be possible and useful to rank three DLs as 'passive', 'somewhat active' and 'active' without assigning absolute values to their positions on the continuum.

Conclusion

Examination of the research record indicates that most user research is conducted on experimental, public DLs supported by research grants and managed by researchers. Future research into digital libraries of all kinds, including commercial digital libraries, will help us refine the framework of attributes presented here. Work will also continue to integrate the different conceptualizations developed by other researchers (Bates, 2002; Borgman, 2000; Fuhr, Hansen, Mabe, Micsik, & Sølvberg, 2001; Saracevic, 1995; Saracevic & Covi, 2000). Future work will also help us determine the utility of this framework as a tool for articulating digital library evaluation goals, objectives and results for both within-DL and between-DL evaluation projects.

References

Bates, M. J. (2002). The cascade of interactions in the digital library interface. *Information Processing & Management, 38*(3), 381-400.

Bishop, A. P., & Star, S. L. (1996). Social informatics of digital library use and infrastructure. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)* (Vol. 31, pp. 301-401). Medford, NJ: Published for the American Society for Information Science (ASIS) by Information Today.

Borgman, C. L. (2000). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press.

Fuhr, N., Hansen, P., Mabe, M., Micsik, A., & Sølvberg, I. (2001). Digital libraries: A generic classification and evaluation scheme. In P. Constantopoulos & I. T. Sølvberg (Eds.), *Research and advanced technology for digital libraries 5th European Conference, ECDL 2001, September 4-9, 2001* (pp. 25-40). Darmstadt, Germany.

Lesk, M. (1997). *Practical Digital Libraries: Books, Bytes, and Bucks*. San Francisco, CA: Morgan Kaufmann.

Lynch, C. (in press). Colliding with the Real World: Heresies and Unexplored Questions about Audience, Economics and Control of Digital Libraries. In A. P. Bishop, B. P. Buttenfield, & N. A. v. House (Eds.), *Digital Library Use: Social Practice in Design and Evaluation* . Cambridge, MA: MIT Press.

NSF. *Digital libraries initiative, phase one (DLI-1)*. National Science Foundation. Available: http://www.dli2.nsf.gov/dlione/ [2001, November 1].

Rice, R. E., McCreadie, M., & Chang, S.-J. L. (2001). *Accessing and browsing information and communication*. Cambridge, MA: MIT Press.

Saracevic, T. (1995). *Evaluation of evaluation in information retrieval.* Paper presented at the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Saracevic, T., & Covi, L. (2000). Challenges for digital library evaluation. In D. H. Kraft (Ed.), *Knowledge Innovations: Celebrating Our Heritage, Designing Our Future. Proceedings of the 63rd Annual Meeting, November 11-16, 2000, Chicago, IL* (pp. 341-350). Washington, D.C.: American Society for Information Science.

Position Paper
On the Assessment of Scalability of Digital Libraries

Christian Mönch
moench@idi.ntnu.no

Abstract

In this position paper an attempt is made to identify the factors that influence scalability of digital libraries. These include not only currently known states, but also assumptions about the future state of certain factors. A cost based measure for scalability is given, which shows how the identified factors may be used to assess the scalability of digital libraries.

1 Introduction

"Digital Libraries are concerned with the creation and management of information resources, the movement of information across global networks and the effective use of this information by a wide range of users" [from the introduction in the first issue of "International Journal of Digital Libraries", Springer, 1997]. To provide their services over a long time, digital libraries have to be scalable to support for future growth in different dimensions, for example in the number of users and/or in the number of stored documents. The assessment of scalability is therefore an important aspect of the assessment of digital libraries. Obviously it is not an easy task to assess a potential that will manifest itself in the future, because the assessment can never be based on presently available informations only. It must, at least to some extent, include assumptions about future development.

This paper outlines an approach to assess the long term scalability of a digital library under these assumptions. For the scope of this paper it is assumed, that the evaluated digital library is effective, i. e. it fulfills the tasks it was designed for (cf. [7]). Therefore scalability will be measured in terms of costs, i. e. the cost of required resources.

In the following section a strategy for the assessment of scalability of digital libraries in an evolving environment is outlined. Examples for the application of this strategy are given in section 3. Section 4 summarizes the result and outlines some of the work that remains to be done.

2 Assessing potential growth

The assessment of long term scalability has two different aspects. The first aspect is concerned with the predictable, or at least analyzable, behavior of a digital library, i. e. the dimensions of growth and the resource usage of the digital library when it grows.

The second aspect is related to changes in the environment of the digital library. Because digital libraries are envisioned to provide their service for years, it is likely that their environment changes during their runtime. So in general change should be expected in a digital library (cf. [4]). This aspect can again be divided into two parts, the expected growth in the different dimension of growth, and the future costs of required resources. Therefore the resource usage should be set in relation to the environment of the digital library. It is obvious, that the future development can only be estimated. This aspects of scalability assessment therefore exposes a certain amount of uncertainty.

2.1 The factors of scalability

The following factors for scalability of a digital library have been identified: the dimensions of growth, the resource requirements of the digital library, the estimated costs of resources, and the assumed growth rates of the growths dimensions. These factors will be outlined in turn in the following.

Scaling dimensions When planing or assessing a digital library it has to be defined, which variables should be scaled, i. e. in which dimensions the digital library should grow. A digital library may grow in the number of stored documents and/or in the number of supported users. While these are the most common scaling variables, there may be others. For example, the number of different document types or the number of different services. The scaling dimensions are represented by a set of variables $\{dim_1, dim_2, \ldots, dim_m\}$.

The dimensions of scaling (not the size of individual variables) belong to the definable factors of scalability. After they are settled it is in general nearly impossible to change them. The reason for this is, that scalability requires the builtin ability to change, which is, due to efficency reasons, usually restricted to the scaling dimensions. Therefore a change of scaling dimensions would require the development and/or employment of a new digital library system, which is almost never viable.

Resource requirements The resource requirements describe how the resource consumption of the digital library is related to the size of the scaling variables. Resources may be, for example, the amount of used storage or the amount of necessary bandwidth to support the operation of the digital library. The resource requirements are predictable magnitudes because they are determined by the architecture of the digital library.

Resources are represented by a set $\{r_1, r_2, \ldots, r_n\}$. Resource requirements are represented by a set of functions $\{rq_{r_1}, rq_{r_2}, \ldots, rq_{r_n}\}$, where rq_{r_i} maps a vector of all scaling variables to the required amount of resource r_i.

Resource costs In order to assess the overall costs of the digital library, the future costs of the resources have to be estimated. This is done by a set of functions $\{rc_{r_1}, rc_{r_2}, \ldots, rc_{r_n}\}$, where rc_{r_i} maps a time value to the assumed cost per time unit of resource r_i at the given time.

Growth rates In order to assess the scalability of a digital library, the expected growth of the scalable dimensions has to be specified. A growth rate is a function describing the size of the scaling variable at a certain time. The growth rates can be given as a set of functions $\{g_{dim_1}, g_{dim_2}, \ldots, g_{dim_m}\}$, where g_{dim_i} maps a time value to the assumed value of the scaling variable dim_i at a given time.

2.2 Assessment of scalability

With the given definitions the overall costs per time unit $c(t)$ of a digital library at a given time t evaluate to:

$$c(t) = \sum_{i=1}^{n} rc_{r_i}(t) * rq_{r_i}(g_{dim_1}(t), g_{dim_2}(t), \ldots, g_{dim_m}(t))$$

The scalability of the digital library depends on the costs per time unit, the operator of the digital library can provide. These costs are denoted by the function tc that maps a time value to costs per time unit. The digital library in question is scalable, if the condition $c(t) \leq tc(t)$ holds for all t.

3 Example for a scalability assessment

In this section the scalability of two infrastructures for digital libraries will be investigated, i. e. an infrastructure for digital libraries called Indigo, and the basic infrastructure of the World Wide Web.

Both infrastructures support the scalability of the number of documents and the number of users. In contrast to the World Wide Web, Indigo has been designed to support the scalability of document types and the scalability of the supported services. This was done by identifying a set of basic operations, the so called orthogonal operations, that can be performed on every document and making the mapping from operations onto processes that perform the

operation explicit in order to support automated modifications. This allows for cheap introduction of new document methods and these can in turn be used to transparently integrate new services into a digital library. Different aspects of Indigo are described in [6] and [5].

The factors of scalability, that are independent of the infrastructures, are described below. A investigation of the specific aspects of the two infrastructures follows in the next subsections. Because no empirical data about the growth rate of growth dimensions is available, and because the growth rate also depends on the individual digital library, a coarse measurement is applied. For the purpose of this discussion it will be distinguished between constant, logarithmic, linear, and exponential resource requirements, resource costs, and growth rates.

Scaling dimensions They scaling dimensions are: the number of stored documents, the number of supported users, the number of newly introduced document types and the number of newly introduced services, i. e. $\{docs, users, newtypes, newservices\}$.

Growth rates The growth rates of documents and users are assumed to be linear, i. e. $g_{docs}(t) := t$, $g_{users}(t) := t$. The rates of introducing new document types and new services are supposed to be logarithmic $g_{newtypes}(t) := ln(t)$, and $g_{newservices}(t) := ln(t)$.

Resource costs The following resources are taken into consideration: Processing time, storage capacity, network traffic volume, and human resources, i. e. $\{cpu, store, traffic, human\}$. For the costs of these resources, the following functions are assumed: $rc_{cpu}(t) := e^{-t}$, $rc_{store}(t) := e^{-t}$, $rc_{traffic}(t) := 1$, and $rc_{human}(t) := t$.

3.1 Indigo

A coarse estimate of the resource requirements in Indigo is given by the following functions:

cpu Indigo uses processing time to perform operations on documents and services. Operations and services are invoked by users and therefore depending on the number of users. The amount of processing time needed to fulfill a service depends on the number of documents: $rq_{cpu}(docs, users, newtypes, newservices) := users(1 + docs)$.

store Indigo stores documents in loosely coupled storage components. The amount of storage needed is depending on the total number of documents: $rq_{store}(docs, users, newtypes, newservices) := docs$. In addition it is assumed, that the number of storage components is: $ln(docs)$.

traffic The volume of the network traffic depends mostly on the number of storage components and on the number of different document types on these storage components. The reason for this is, that mobile service documents are transmitted to every storage component in order to perform a service. It is assumed that the number of different document types on a storage component is: $ln(docs)$. In addition the traffic depends on the number of invoked document operations. Since the number of invoked document operations and services is depending on the number of users, one gets: $rq_{traffic}(docs, users, newtypes, newservices) := users * (1 + ln(docs) * ln(docs))$.

human The required human resources are depending on the total amount of administrative work needed to run the infrastructure, which is depending on the number of storage components: $rq_{human}(docs, users, newtypes, newservices) := ln(docs)$.

So the overall costs per time unit in Indigo $c_{indigo}(t)$ at time t are:

$$c_{indigo}(t) = (e^{-t} * (t + t^2)) + (e^{-t} * t) + (t + t * ln(t)) + (t * ln(t)^2) \Leftrightarrow$$

$$t^2 * e^{-t} + 2t * e^{-t} + t * ln(t)^2 + t * ln(t) + t$$

This function is bounded by the function $t * ln(t)^2$, so it grows faster than linear, but slower than quadratic.

3.2 WWW-based Digital Libraries

The scalability assessment is now performed for a digital library that uses the World Wide Web as base technology, i. e. library side web server with some kind of scripting mechanism. The resource requirements are as follows:

cpu The cpu usage in the WWW does also depend on the number of users and the number of documents: $rq_{cpu}(docs, users, newtypes, newservices) := users(1 + docs)$.

store It is assumed that the amount of storage needed in the WWW-based approach is depending on the total number of documents: $rq_{store}(docs, users, newtypes, newservices) := docs$. Again it is assumed, that the number of distinct WWW-servers is: $ln(docs)$.

traffic The volume of the network traffic depends on the number of invoked HTTP-request and service calls, which in turn depend on the number of users. The resulting traffic is supposed to be constant for the presentation of a document and depending on the number of documents for the result of a search request: $rq_{traffic}(docs, users, newtypes, newservices) := users * (1 + ln(docs))$.

human The required human resources are also depending on the administrative work, which is also assumed to grow logarithmicaly with the number of documents. In addition, and in contrast to the Indigo, human resources are needed to install methods for new document types on every storage component. The number of storage components is supposed to increase logarithmically to the number of documents. Therefore the requirements for the resource $human$ are: $rq_{human}(docs, users, newtypes, newservices) := ln(docs) * (1 + newtypes * newservices)$.

So the overall costs per time unit in a WWW-based digital library are given by:

$$c_{www}(t) = (e^{-t} * (t + t^2)) + (e^{-t} * t) + (t + t * ln(t)) + (t * ln(t)^3) \Leftrightarrow$$

$$t^2 * e^{-t} + 2t * e^{-t} + t * ln(t)^3 + t * ln(t) + t$$

The costs for the WWW-based implementation of a digital library with the desired properties grows faster than the costs for the Indigo-based implementation because of the requirements for the resource $human$. Manual work is needed to include the handling of new document types into the typical script-based approach. This indicates that a WWW-based infrastructure is expensive, if the implemented digital library should be highly scalable in the dimension of document types and/or services.

4 Conclusion

This position paper presented some thoughts on the assessment of long term scalability of digital libraries. A number of factors that are relevant for the assessment of scalability have been identified. Beside the known factors *scaling dimensions* and *resource requirements* the uncertain factors *resource costs* and *growth rate* have been identified. It was shown how a cost-based measure could be derived from these factors.

The scalability of every digital library is influenced by the development of certain environmental magnitudes. Even if they can not be predicted precisely, they can at least be used to clearly state the assumptions that have been made during the design of a digital library. The explicitly stated assumptions may also be used to drive the design process for digital libraries. The necessity to explicitly state the assumed scaling behavior has also been recognized in other areas, e. g. in parallel computing. In [2] a scaling path is described which defines a growth rate and a resource requirement.

It should be noted, that some basic assumptions have not been mentioned. For example, a necessary precondition for scalability is scalability of the technological infrastructure, the digital library is build on. If this is not given, the digital library has to be refitted to a new infrastructure. The first implementation of NCSTRL is an example for such a case (cf. [3].)

4.1 Future work

The paper provides two examples which illustrate the application of the cost-based measure. Due to missing empirical data the cost functions are only showing asymptotic behavior. This will not be sufficient to judge the short term or medium term, i. e. five to ten years, costs of a digital library. To attempt a more detailed prediction, empirical data about the development of resource costs, and the resource requirements has to be gathered. For example, a yearly increase in the number of documents of 25 percent could be modeled by the function $g_{docs}(t) := 1.25^{t/12} * docs_0$, where t is months from now, and $docs_0$ is the current number of documents. Similarly the costs for processing power could be modeled to halve every 18 months, assuming that Moore's law holds, that doubled transistor number means doubled processing power, and that processor prices are remaining constant. The resulting cost function for processing power would be: $rc_{cpu}(t) := 0.5^{t/18} * cpu_0$, where t is month from now and cpu_0 refers to the current cost of processing power.

The prediction of future states is likely to contain some errors. These errors will accumulate, as the time span to be assessed respectively predicted grows. If one is uncertain about a future state, and therefore likely to make error prone predictions, methods for reasoning under uncertainty as they are used, for example, in software engineering (see [1]), could be taken into account, in order to augment scalability predictions with an indication of their likelihood.

References

[1] Barry W. Boehm. *Software Engineering Economics*. Prentice Hall, 1981.

[2] Prasad Jogalekar and Murray Woodside. Evaluating the Scalability of Distributed Systems. *IEEE Transactions on Parallel and Distributed Systems*, 11(6):589–603, June 2000.

[3] Kurt Maly, Mohammad Zubair, Hesham Anan, Dun Tan, and Yunchuan Zhang. Scalable Digital Libraries Based on NCSTRL/Dienst. In José Borbinha and Thomas Baker, editors, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2000*, number 1923 in Lecture Notes in Computer Science, pages 168–179, Lissabon, Portugal, September 2000. Springer-Verlag.

[4] Alexa T. McCray and Marie E. Gallagher. Principles for Digital Library Development. *Communications of the ACM*, 44(5):49–54, May 2001.

[5] Christian Mönch. INDIGO — An Approach to Infrastructures for Digital Libraries. In José Borbinha and Thomas Baker, editors, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2000*, number 1923 in Lecture Notes in Computer Science, pages 158–167, Lissabon, Portugal, September 2000. Springer-Verlag.

[6] Christian Mönch and Oswald Drobnik. Integrating New Document Types into Digital Libraries. In *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries IEEE ADL '98*, pages 56–65. IEEE, April 1998.

[7] Tefko Saracevic and Lisa Covi. Challenges for digital library evaluation. In *Proceedings of the American Society for Information Science*, volume 37, pages 341–350, 2000.

A quality management approach to the evaluation of digital library services

Jillian R. Griffiths
Research Fellow
CERLIM (Centre for Research in Library and Information Management)
Department of Information and Communications
Manchester Metropolitan University, UK
e-mail: j.r.griffiths@mmu.ac.uk

Shelagh Fisher
Reader
Department of Information and Communications
Manchester Metropolitan University, UK
e-mail: s.m.fisher@mmu.ac.uk

Abstract

This paper examines the use of a Quality Management approach in the evaluation of digital services, and presents indicative results of this approach following user testing in the UK of the Joint Information Systems Committee's Distributed National Electronic Resource (JISC's DNER). The Study was undertaken as part of the EDNER (Formative evaluation of the DNER) project, a three year project funded by JISC which commenced in July 2000. Test criteria draw upon 'Quality Attributes' which were first posited by Garvin in 1987 in the context of retail and service industries and subsequently applied to information services by Brophy in 1998. This paper reports on the use of the quality attributes in the EDNER project to evaluate JISC's digital information services and projects from an end-user perspective.

Introduction

The Distributed National Electronic Resource (DNER, www.jisc.ac.uk/dner) is being developed by the Joint Information Systems Committee (JISC), which is a strategic advisory committee working on behalf of the funding bodies for Higher and Further Education in England, Scotland, Wales and Northern Ireland. The DNER is the working title for the concept of an electronic resource which appears seamless to the user. The intention underpinning the DNER is that staff and students in higher (HE) and further (FE) education will be able to access resources effectively and efficiently through intuitive and customised interfaces, unlike current electronic services where the user needs to know the name of the service provider (JISC, 2001).

The EDNER Project (Formative evaluation of the DNER, www.cerlim.ac.uk/edner) was funded by JISC to undertake formative evaluation of the developing DNER over a three year period from 2000 to 2003. The Project is being led by the Centre for Research in Library and Information Management (CERLIM) at Manchester Metropolitan University in partnership with the Centre for Studies in Advanced Learning Technology (CSALT, www.comp.lancs.ac.uk/csalt/) at Lancaster University.

This paper examines the use of a Quality Management approach in the evaluation of digital services, and presents indicative results of this approach following user testing in the UK of the JISC's DNER. Test criteria draw upon Quality Attributes as applied to information services. This paper reports on the use of the Quality Attributes in the EDNER project to evaluate JISC's digital information services and projects from an end-user perspective.

Quality Attributes

The approach used here maps to many of the quality assurance approaches which government in the UK is sponsoring – for example in UK public libraries there is focus on 'Best Value'; in European business circles, the talk is of 'business excellence'; and the European Foundation for Quality Management has re-titled its annual award as

'The European Award for Business Excellence'. A key aspect of quality assurance is its emphasis on the satisfaction of all the stakeholders.

Garvin (1987) identified eight attributes that can be used to evaluate a variety of services. These were adapted and extended by Brophy (1998) to apply to information and library services as: 1) Performance, 2) Conformance 3) Features, 4) Reliability, 5) Durability, 6) Currency, 7) Serviceability, 8) Aesthetics and image, 9) Perceived quality and, 10) Usability.

Performance is concerned with establishing confirmation that a service meets its most basic requirement. These are the primary operating features of the product or service.

With **Conformance** the question is whether the product or service meets the agreed standard. This may be a national or international standard or locally determined service standard. The standards themselves, however they are devised, must of course relate to customer requirements. It may be a question of whether they have addressed interoperability issues and how they utilise emerging standards such as XML, RDF, Dublin Core, Z39.50 etc.

Features are the secondary operating attributes, which add to a product or service in the user's eyes but are not essential to it. It is not always easy to distinguish 'performance' characteristics from 'features', especially as what is essential to one customer may be an optional extra to another, and there is a tendency for 'features' to become 'performance' over time.

Users place high value on the **Reliability** of a product or service. For products this usually means that they perform as expected (or better). For information services, a major issue is usually availability of the service. Therefore broken links, unreliability and slowness in speed of response can have a detrimental affect on a user's perception of a service.

Garvin uses the term **Durability**, defined as 'the amount of use the product will provide before it deteriorates to the point where replacement or discard is preferable to repair'. In the case of information services this will relate to the sustainability of the service over a period of time. In simple terms, will the service still be in existence in three or five years?

For most users of information services an important issue is the **Currency** of information, i.e. how up to date the information provided is when it is retrieved.

Serviceability relates to when things go wrong, how easy will it be to put them right? How quickly can they be repaired? How much inconvenience will be caused to the user, and how much cost? For users of an electronic information service this may translate to the level of help available to them at the time of the search. So the availability of instructions and prompts throughout, context sensitive help and usefulness of help will be important.

Whilst **Aesthetics and Image** is a highly subjective area, it is of prime importance to users. In electronic environments it brings in the whole debate about what constitutes good design. In a web environment, the design of the home page may be the basis for user selection of services, and this may have little to do with actual functionality. You may have a great information service behind that home page, but do the users ever find it?

Perceived Quality is one of the most interesting of attributes because it recognises that all users make their judgments on incomplete information. They do not carry out detailed surveys of 'hit rates' or examine the rival systems' performance in retrieving a systematic sample of records. Most users do not read the service's mission statement or service standards and do their best to by-pass the instructions pages. Users will quickly come to a judgment about the service based on the reputation of the service among their colleagues and acquaintances, their preconceptions and their instant reactions to it.

The addition of **Usability** as an attribute is important in any user-centred evaluation. User-centred models are much more helpful when personal preferences and requirements are factored in – so, for example, usability to a blind person may mean something quite different to usability to a sighted person.

Methodology

The aim of the EDNER study was to evaluate the quality of DNER services according to a range of defined criteria (Quality Attributes). This was achieved by firstly establishing the Quality Attributes, with appropriate revisions and adaptations for its use in this context. Test searches were then designed (one for each of the services to be used by the participants, fifteen in total). These searches were designed so that they would be of sufficient complexity to challenge the user without being impossible for them to answer. Participants were recruited via Manchester Metropolitan University's Student Union Job Shop. Twenty-seven students from a wide course range participated. Each student was paid for his or her participation. One third of the sample consisted of students from the Department of Information and Communications and were studying for an Information and Library management degree, the remaining two thirds of the sample were studying a wide variety of subjects and all were at various stages of their course. No restrictions were placed on them to have computer, searching or Internet experience. Testing was conducted in a controlled environment based within the Department of Information and Communications, Manchester Metropolitan University. Each participant searched for the fifteen test queries and completed questionnaires for each task undertaken. Data gathered via the questionnaires was analysed in two ways, 1) quantitative data was analysed using SPSS Statistical Package for the Social Sciences), and 2) open response question data was analysed using qualitative techniques.

Table 1 shows the measures which were developed for each of the Quality Attributes. The majority of these measures came from existing (and generally accepted) measures. Measures used focussed on a user-centred approach.

Quality Attribute	Measure
Performance	• Satisfaction that required information was found • Satisfaction with ranking order of retrieved items
Conformance	Not evaluated by users
Features	• Search option/s used • Features particularly liked
Reliability	• Any dead links found • Impact of dead links on judgment of service • Satisfaction with speed of response
Durability	Not evaluated by users
Currency	• Information retrieved by the service up-to-date
Serviceability	• Instructions and prompts helpful • Use of Help • Helpfulness of Help
Aesthetics	• Satisfaction with interface and presentation of features • Familiarity with interface/elements of the interface • Ease of understanding of retrieved item list
Perceived quality	• Rate quality of service and information retrieved
Usability	• User friendliness of service • How easy to remember which features to use • Satisfaction with facility to input query • Satisfaction with facility to modify query

Table 1 Quality Attributes and measures

Sample results

Figure 1 represents results of three of the quality attributes (Performance, Usability and Aesthetics) across five of the services evaluated during the testing. Results are also given for the Perceived quality attribute **pre** testing (before the participants used the services in the test) and **post** testing (after participants used the service). Where participants

were aware of the service prior to testing they were asked to indicate their perception of the quality of the service and the information retrieved.

The final measure presented here is that of Overall satisfaction. On completion of each task participants were asked to give an overall rating indicating their satisfaction for each service, taking into account the results retrieved, ease of retrieval, satisfaction with the interface etc. Likert type scales were used for each measure and Figure 1 combines those responses where participants indicated good or excellent ratings.

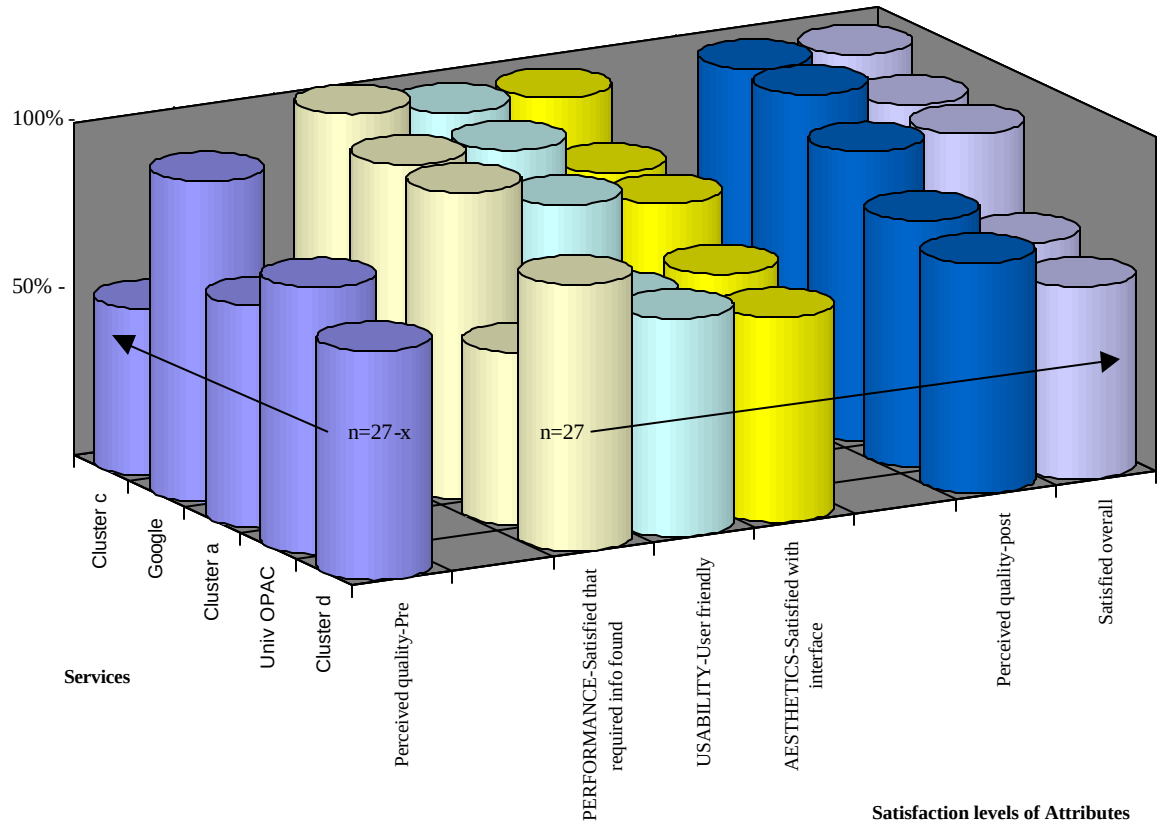


Figure 1 User satisfaction with services by attribute

Results show varying degrees of satisfaction across each of the services and each of the Attributes. On Service D, Perceived quality pre and post searching remained static (and relatively low), despite high levels of satisfaction with Performance. In conjunction with this performance score, satisfaction with Usability and Aesthetics were lower. Overall Satisfaction was also relatively low. This seems to indicate that users' perceptions of quality are driven by factors other than Performance of a system. It also raises interesting questions as to how fixed preconceptions about quality may affect the results of the evaluation of a system or service.

On the University OPAC post-search Perceived Quality dropped only very slightly (2%), despite low levels of satisfaction with Performance. Satisfaction with Usability and Aesthetics was slightly higher than that of Performance. This again may indicate that factors (or Quality Attributes) such as User Friendliness and Interface Design may be at least as important as evaluation criteria as Performance.

Users expressed an increase in post-search Perceived Quality on Services A and C, coupled with high levels of Satisfaction Overall and across each of the Attributes. In this instance actual use of the service appears to have changed the preconceptions of the users.

Results for Google were high across all the measures used in the test. However, satisfaction with Aesthetics was lower than satisfaction with Aesthetics on Service C, and had lower Overall satisfaction ratings.

In each instance Satisfaction Overall corresponded very closely with post Perceived Quality.

Conclusions

Use of the Quality Attributes as evaluation criteria allows investigation of what happens in between perceptions of quality of service before and after use of the service. This allows for improvement of services by targeting areas that have scored lower. Therefore, this approach allows service providers and developers to identify specific areas for improvement. In IR terms, performance would traditionally be measured by recall and precision, and in a Web environment it may be measured by user satisfaction (Johnson et al 2001). These results are early indicators from work in progress but seem to demonstrate that other measures play an important role in user evaluation. Using a Quality Management approach, we can demonstrate that users' preconceptions play a major role in their evaluation of a service and can be are hard to shift – an example is Service D where Pre and Post perceived quality did not change despite the fact that users were satisfied that required information was found. However, where a service performs well across *all* attributes user perceptions can change through use of the service (for example Service C). This raises interesting questions about how services are developed, how users are trained and how services are marketed.

Use of the Quality Attributes allows for a holistic approach to the evaluation of digital library services which can take into account different stakeholder perspectives and different services. This research utilised a user-centred approach, but could be tailored to different contexts and stakeholders.

References

Brophy, P. (1998) It may be electronic but is it any good? Measuring the performance of electronic services. In: *Robots to Knowbots: the wider automation agenda: Proceedings of the Victorian Association for Library Automation 9th Biennial Conference*, January 28-30 1998 Melbourne: VALA, pp. 217-230.

Garvin, D.A. (1987). Competing on the eight dimensions of quality. *Harvard Business Revie*w: pp. 101-109.

JISC (2001) *Five-Year Strategy 2001-05 Executive Summary*
http://www.jisc.ac.uk/pub01/strat_01_05/exec.html#s3

Johnson, F.C., Griffiths, J.R. and Hartley, R.J. (2001) *DEVISE: A framework for the evaluation of Internet search engines* Resource: The Council for Museums, Archives and Libraries (Library and Information Commission Research Report 100) Available at
http://www.cerlim.ac.uk/projects/devise.htm

Plutchak, T.S. (1989). On the satisfied and inept user. *Medical Reference Services Quarterly* 8 (1): pp 45-48.

DELOS Workshop on evaluation of digital libraries, Budapest, June 2002

TEL: The European Library
Developing a user-oriented digital library evaluation for an IST Accompanying Measure

by Linda Banwell
Information Management Research Institute, School of Information Studies, Northumbria University, Newcastle upon Tyne, UK

Summary of themes and issues to be covered in the paper

The paper is about the development of an evaluation framework for the TEL (The European Library) project. The project is in progress, and the paper reports on the research and issues raised to date. The aim of the paper is to inform, and provoke discussion at the Workshop. Themes and issues raised by TEL, can be summarised as follows:

- Project evaluation in the 'EU funded project' context: importance and scale of such activity,
- Project vs. service evaluation in the context of an EU Accompanying Measure (AM): TEL as an AM stops short of becoming the full service it would have been as a full RTD project,
- Appropriate methodology for user-oriented evaluations: qualitative; developmental framework,
- Practical considerations: cultural and language differences; data collection problems,
- Transferability of methods and findings from TEL evaluation to other projects and contexts: sharing problems and solutions, and disseminating widely.

About TEL

The TEL (The European Library) project began in February, 2001. It is a 30 month project, funded by the European Commission as part of its Fifth Framework Programme for research. It aims to set up a co-operative framework for access to the major national, mainly digital, collections in European national libraries. TEL is funded as an Accompanying Measure, designed to support the work of the IST (Information Society Technologies) programme on the development of access to cultural and scientific knowledge. TEL will stop short of becoming a live service during the lifetime of the project, and is focused on ensuring co-operative and concerted approaches to technical and business issues associated with large scale content development. It will lay the policy and technical groundwork towards a pan European digital library based on distributed digital collections, and providing seamless access to the digital resources of major European national libraries. TEL has 8 national library partners: Finland, Germany, Italy, the Netherlands, Portugal, Slovenia, Switzerland and the United Kingdom. It is also seeking to encourage the participation of all European national libraries, in due course.

The project activity is being pursued through six Workpackages, which are all progressing in parallel, and will come together at the end of the project in 2003. All the main Workpackages (1-4) have completed scoping studies, typically containing literature review and survey work, and are now into their main phase of development work. The Workpackages are focusing on:

Workpackage 1 – Publisher relations
The consortium is working with publishers of electronic materials and publisher organisations to establish co-operative approaches to business, licensing and copyright matters.

Workpackage 2 – Business plans and models
Designed to maximise the benefits of co-operation through the development of joint or individual business plans and models ready for implementation in the operational phase of TEL.

Workpackage 3 – Metadata development

Aims to develop a concerted best practice approach to metadata standards and schemas to support access to digital material. The agreed approaches will be tested in Workpackage 4.

Workpackage 4 – Interoperability testbeds
This Workpackage is carrying out preparatory work prior to the development of the operational service. It will produce a functional specification, and benchmarks will be defined from the outputs of Workpackages1-3. There are 2 testbeds under development in parallel:
1. focused on distributed searching using Z39.50
2. focused on aggregated, centralised searching using XML

Interoperability will be achieved through the centralised TEL portal. Development work will be complete by September, 2002, to be followed by a period of functionality testing, which will include scalability of access and multi-lingual capability. Interoperability is seen as key to demonstrating the proof of concept of TEL within its user community.

Workpackage 5 – Dissemination
Reports are available on the TEL website at http://www.europeanlibrary.org
The TEL Milestone conference held in Frankfurt in April, 2002, was the first project wide dissemination event, which aimed to inform widely about TEL. It was attended by around 120 delegates from most European countries.

Workpackage 6 – Management aspects
These include the evaluation activity.

Developing an evaluation framework for TEL

Included in the project plan is the objective of developing an evaluation framework for TEL, as a benchmarking baseline to be implemented by the partners once TEL becomes a live system. TEL is being evaluated at 3 levels:

1. Completion of Workpackage deliverables, to be monitored by the TEL Project Manager.
2. Overall project performance in relation to the IST programme priorities.
3. Establishing user viewpoints as the basis of an evaluation framework to be used as a benchmarking baseline for the eventual live TEL service.

The second and third levels of evaluation are being undertaken by an independent evaluator, and it is with these that this paper is primarily concerned.

The broad context of TEL as an EU funded project provides the basis for the evaluation framework. There are several aspects, which provide a starting point:

- the *Information Society Programme themes*, for example, the "informed citizen" is a prominent theme – do TEL partners have a profile in this context, and what is their approach to serving the general public? Has community added value and contributions to community social objectives been achieved? Have community economic development and dissemination criteria been met?
- the requirements associated with being an *Accompanying Measure*: the project evaluation, which would be suited to a full RTD project, is not appropriate for TEL, which will ultimately need to incorporate the evaluation of an on-going service. The evaluation of the TEL project phase therefore needs to lay the foundation for an eventual service evaluation.
- at *partner level*, to discuss the added value to be had from participating in TEL for the library and its users, for example, reviewing the extent to which TEL is encouraging partners to develop resources that, in co-operation, are greater than any would have on its own
- *common questions* to be asked when evaluating a europroject are about the short to medium term outputs of the project, the learning gained and the extent of its diffusion leading to the establishment of new practices which might act as the catalyst for further change

A variety of activity is building up different layers of information in support of the evaluation. This is the broad pattern, which has been followed in TEL. In the first year of the project, meetings were held within the project team in order to develop the framework for the evaluation. These were followed by broader based meetings between partners in the project, and meetings with the wider stakeholder group, at the TEL Milestone conference held in Frankfurt in April, 2002. Dissemination about the project is on-going.

TEL evaluation: methods

There will be two outputs from the evaluation and monitoring activity:

- The overall, *summative project performance review*. This will be end-of-project activity, which will include synthesis of Workpackage level evaluations and other documentation generated in the project, and of data collected direct from partners and partner communities.
- *Formative, monitoring and evaluation activity* designed to establish and monitor TEL user viewpoints, link back in to the evaluation of specific Workpackages in an on-going way, and inform the development of a benchmarking baseline for TEL.

The TEL evaluation is essentially user centred, where qualitative data are as important as quantitative data. It is multifaceted and multilayered, with data triangulated to enhance its trustworthiness and objectivity. Formative evaluation gives a longitudinal character to the evaluation, whereby it becomes problem solving, feeding back into the project. The approach aims to be holistic, naturalistic. It is taking place in an uncertain environment: the future shape of digital libraries is unknown, meaning that the evaluation must provide the partners with a flexible and adaptable tool for their future use. It is evaluation with a research focus - it seeks both to explain the specific digital library, and relates it to wider issues. It is also developmental, in that the results of the evaluation are fed back in to the project, as it progresses. The approach builds on existing work, by the author (Banwell, 2000; Banwell, Day and Ray, 1999; Banwell and Gannon-Leary, 2001) and others, such as (Kelleher, Sommerlad and Stern, 1996; Marchionini, 2000; Saracevic, 2000).

The evaluation process is progressing through taking the following steps:

- Establish a written *evaluation plan*, taking into account EU project evaluation requirements, and building on best practice in the field. The plan was delivered in September, 2001, and included in documentation for the first six monthly review by the EU. The feedback received stressed the importance to TEL of finding out in detail about its potential users, thereby reinforcing the approach being taken in the TEL evaluation.
- *Documentary analysis* of the detail in the scoping exercises carried out by all Workpackages, provided a detailed starting point for the evaluation. The market research activity in WP2 has been especially useful in providing information on user contexts, views and service evaluation criteria.
- *User panels* were to be established by all partner libraries to represent as many views as possible from within the national user communities, and as an important mechanism for on-going contact with users and potential users of TEL
- *Formative evaluation* is being undertaken through structured six monthly contact with users and potential users, with contact made electronically, or through face-to-face contact at Workpackage meetings, and at conferences (such as the TEL Milestone conference at Frankfurt in April, 2002) and workshops. Questions will seek to identify barriers and enablers for the development of TEL in the broad areas of content (issues here will be accessibility, accuracy, richness), technical infrastructure (e.g. speed, ease of use), and user perceptions of the role, importance and potential impact of TEL to users as individuals, and nationally. Shared understandings are being sought through investigation of variations at national and individual levels.
- Formative evaluation activity is *reported* through the preparation of six monthly reports for the TEL project manager. It will also feed into the summative evaluation report at the end of the project, with appropriate dissemination.

TEL evaluation: the developing picture

Scoping studies

The TEL scoping studies carried out to date show that the current situation with regard to users and use of digital resources in TEL partner libraries is largely a continuation of the pre-digital situation. The current emphasis of collections and services is on national language, literature and culture, used largely by higher level students, academics and some personal researchers/authors. Most partners have few or no statistics about their current users and usage. Where data are available they are for traditional reading room registrations. No partners have data on their internet users. There is a problem with the comparability of even what data are available – definitions and the recording of statistics vary between partners. Nevertheless, variations in practice between partners are suggested, which provide an illuminating baseline for the project. For example, only one of the eight national library partners records that it keeps sound recordings, illustrations, and photographs as a digital

resource, and only three of the eight keep newspapers and dissertations. Indeed, the greatest consensus between the eight partners is that five of them record that they do, or plan to, keep scientific journal articles in digital format.

Questionnaire data

The TEL Milestone conference in April, 2002, was attended by around 120 delegates from the TEL partner national libraries, from European national libraries not yet part of the TEL project, and from the wider potential user community for TEL. All delegates were asked to complete a short questionnaire in order to establish their current information behaviour, especially in relation to digital resources, and for their views on the added value that TEL might bring them.

26 completed questionnaires were received, a disappointing response rate of just over 20%. A brief digest of the responses is included, as follows:

- Most respondents used digital information sources "very often", with greatest use being made of the Internet, followed by CD-ROMs and e-journals/newspapers.
- Only a handful of respondents used digital resources provided by their national libraries. Most were "fairly satisfied" with the digital resources they used.
- Half the respondents found cost to be a major barrier to their use of digital resources. Access and content were not seen as barriers.
- Researchers, academics and students were seen as the main potential users of TEL.
- The major problems envisaged for TEL were:
  - variations in standards - 7 respondents
  - cost and financing - 6 respondents
  - language issues - 3 respondents
- The added value of TEL for individuals was seen by 7 respondents as "TEL as a one-stop shop". The added value of TEL to the national libraries was generally seen as being in the broad area of co-operation and sharing.
- 13 respondents agreed strongly and 6 agreed a little that it is the national libraries' role to deliver the EU concept of "the informed citizen".
- 7 respondents agreed strongly and 12 agreed a little that TEL would succeed in its aim of being the Digital European Library.

The findings from the questionnaire are interesting, but sparse. They may be indicative of more widely held views in the potential user community for TEL. More data are needed if a sound basis for the user-based evaluation is to be obtained. 25 of the respondents, from 7 countries, have offered themselves as participants in further user work, and several other contacts were also initiated. These names will be added to those of the user panels already recruited.

Evaluation outcomes

By the end of the project phase of TEL, partner libraries and their users will have been asked questions to elicit responses, which map on to EU priority concerns. The overall project review will be informed by the answers to questions, such as:

- To what extent has a centralised management process been created for the integrated process of search, locate, order, receive and pay?
- To what extent have centrally negotiated licences for the delivery of material been achieved?
- How much research has happened on non-native language needs?
- Does the TEL business plan make recommendations, which would add value through TEL? Does it indicate the issues of transition from project to operational service? Does it have an outline marketing plan?
- To what extent have common prices and performance measures been agreed between the partners? Has the move towards including qualitative benchmarking in institutions been accommodated?
- Has a strategy to address non-use been developed in each partner library?

TEL is a non-research action, which is about widening participation in the cultural agenda. It is strategically placed, spanning the divide between research and operational service, and between Frameworks 5 and 6. It is

being seen as an enabler, being based on a partnership between national libraries. It should underpin integration, where a key theme is about re-defining national library relationships, including with publishers.

In her Keynote speech at the TEL Milestone conference, Pat Manson, from the European Commission, outlined six core issues to be addressed in the cultural agenda, and which become evaluation criteria for TEL in the widest sense. These issues are content, interoperability, dialogue and alliances, sustainability, quality and dependability, and scaling up (Manson, 2002).

Concluding remarks and points for discussion

A number of issues for resolution have already been raised in relation to the TEL evaluation:

- *About the users* - who and what are they? How to access them in a comprehensive and representative way? Responses from partner countries have varied widely, and have generally been very low in number.
- *About culture, language and meaning.* Making allowance in an evaluation for such wide variations between partner customs and practices, can make comparisons difficult.
- *About user panels* - setting up and establishing communication with user panels in the way originally intended, has thus far not happened in some partner libraries, meaning that the voice of at least part of the TEL user community is not yet being heard. Identifying the wider user constituency is itself a difficulty.
- *About formative evaluation* - it occurs during the life of a project and needs to be fed in, in a meaningful way, which is complex.
- *About evaluating digital libraries* - they are a moving endpoint, making the development of benchmarks difficult at the present time. Appropriate indicators need to be the subject of widespread discussion and investigation in the sector.
- *About moving forward on a very short timescale* - the project lasts 30 months only, is large and complex.
- *About TEL evaluation being very small scale* - it represents 8 weeks effort over the whole of the project, making it a problem to collect enough and the right data in a short time.
- *About TEL evaluation methods* - the strategy is to use multiple research methods in order to collect as broad a range of perspectives and date as is possible in the timescale. It is hoped that the evaluation will succeed in being soundly based, and that it will deliver the required user-based view to help shape the future TEL service.
- *About evaluating an Accompanying Measure* rather than an RTD project. The EU has stressed the importance of the involvement of users. But it is difficult to engage potential users of a service, which will not be operational during the lifetime of the project.
- *About the evaluation type* - is it service or project evaluation? The aim and nature of the evaluation differs between the two types. For TEL the evaluation is of both types: summative for the project and formative for the proposed service, making it a complex activity within its limited budget.
- *Genuine user-based evaluation is a low priority in EU projects* - evaluation is generally seen as being a rather mechanistic, tick box, checklist type of activity. A more broadly user-based evaluation is unexpected and therefore viewed with scepticism by some.

Despite the difficulties raised in the points listed above it is hoped that by doing and disseminating the user-based evaluation activity in TEL, it will thereby be possible to generally raise the profile of user evaluation in europrojects. But, as in all supranational projects, it is very difficult and it should be larger scale and of higher priority for the EU, if their community-wide objectives are to be met. It is hard, immersed at the present time in doing the evaluation, not to see the activity merely in terms of barriers, but it is hoped that it will ultimately succeed in laying a user-based foundation for the future service of TEL.

References

Banwell, L. (2000) Evaluating information services. In: Managing knowledge in health services, edited by Andrew Booth and Graham Walton, Library Association, 2000, pp. 173-181

Banwell, L., Day, J.M. and Ray, K. (1999) Managing organisational change in the hybrid library. Report to UKOLN.

Banwell, Linda and Gannon-Leary, Pat (2001) Meaningful measures for individuals' realities: evidence from the JUBILEE project 4th. International conference on performance measurement in libraries and information services, Pittsburgh, 12-16 August, 2001

Kelleher, John, Sommerlad, Elizabeth and Stern, Elliot (1996) Evaluation of the Electronic Libraries Programme. Guidelines for eLib project evaluation. http://www.ukoln.ac.uk/services/elib/papers/tavistock/evaluation-guide/intro.html [accessed: 21/5/02]
Manson, Pat (2002). Digital culture in the knowledge society. Keynote speech. TEL Milestone conference, Frankfurt, 29-30 April, 2002 [to be published]
Marchionini, Gary (2000) Evaluating digital libraries: a longitudinal and multifaceted view. Library Trends, 49(2), Fall 2000, pp. 304-333
Saracevic, Tefko (2000) Digital library evaluation: toward an evolution of concepts. Library Trends, 49(2), Fall 2000, pp. 350-369
TEL website: http://www.europeanlibrary.org [accessed 21/5/02]

About the author

Dr. Linda Banwell is Director of the Information Management Research Institute and a Principal Lecturer in the School of Information Studies at Northumbria University, Newcastle upon Tyne, England. By background she is a Chartered Librarian, with a doctorate in Computer Science. She specialises in teaching research methods and undertaking user focused research projects, and is the evaluator for the TEL project.

III REPORT OF BREAKOUT GROUPS

Report of Breakout Group on Evaluation in Context

Chair of group:

Nicholas J. Belkin

Members of group:

Linda Banwell, Ann Bishop, Jill Griffiths, Preben Hansen, Laszlo Kovacs, Bob Sandusky, Seppo Saari, Stephan Schneider

Initial task: What is context?

The group identified three major facets of *context* of a DL, as follows:

- Goals and tasks
  - Goals (of the library, organization, users)
  - Tasks (both work tasks and information interaction tasks)
  - Workflow
  - Domain
- Socio-cultural milieu
  - Culture
  - Social practice
  - Politics
  - Language
- Environment
  - Physical
  - Infrastructure
  - Human and fiscal resources
  - Other knowledge resources (outside the DL)

Second task: What are the major research issues/questions with respect to taking account of context in evaluation of DLs?

The group identified five major classes of research questions having to do with context and the evaluation of digital libraries. In the list of questions, below, *X* stands for any of the facets of context (or the isolates within the facets) identified above.

1. How can *X* be usefully classified (where useful is with respect to its applicability to problems of evaluation)?
2. What is the relationship between *X* and evaluation?
3. How can *X* be represented?
4. What is the unit of analysis for *X*?
5. Can evaluation be independent of *X*?

Third task: Methods for investigating the identified research questions.

The group agreed that investigation of context does not imply the use of any particular type of method. Rather, methods should be chosen appropriate to the particular questions and goals of investigation. We did not do a complete inventory of methods appropriate to each of the questions identified above, but did suggest some possible methods for the first two, as follows.

1. How can *X* be usefully classified?

- Ethnographies of situations in which DLs might be implemented. “Might be” is stressed, since this would be most usefully done before the DL exists.
- Synthesis of existing results. Many studies have already been done of such situations, with different characteristics identified as important, and different ways to classify them. A synthesis, or meta-analysis, would be extremely useful.

2. What is the relationship between *X* and evaluation?
   - Hypothesize and test. That is, experimental methods can be appropriate, if sufficient theory is available to generate hypotheses about this relationship.
   - Implement and explore. That is, make a prototype, let people use it within the context, and learn about the relationship from their experience.
   - Operationalize and measure. That is, make concrete some aspect of context, and measure its effect in either operational or experimental system.

Final task: Identify important themes or research questions for research on evaluation of DLs in general.

This task was understood by the group to be a preliminary and broad identification of research areas or projects which need to be investigated and supported in order to establish a firm basis for evaluation of DLs. We also thought it important that the concept of a DL be somewhat generalized, to that of *Complex Networked Information Systems* (CNISs), of which DLs as currently understood are an example. We identified the following as significantly important projects/questions.

- Develop toolkits for evaluation of CNISs (with respect to contextual issues, *inter alia*). These can be based on:
  - Meta-analysis of existing studies (including what has been the effect of evaluation), in concert with, or followed by,
  - Collaborative efforts to compare and evaluate techniques for evaluating CNISs (including multiple evaluations of the same system), which lead to
  - Development of standard techniques, methods, measures for CNIS evaluation, made available to the entire CNIS community.
- Develop a testbed of user interactions in CNISs. Such a testbed could be constructed in a distributed manner, by different groups reporting the results of separate studies to a central agency, which would maintain the testbed. This would almost certainly require developing a set of standards for data collection. The advantage to contributing to such a test bed would be access to all of the other records of user interactions. This is an important activity, since it would allow investigation of a variety of issues by many different groups, without their having to collect data *de novo*.
- Research on how to relate the evaluation of different aspects of CNISs to one another. At the present time, evaluation of different aspects of CNISs proceeds largely without reference to the results of evaluations of other aspects. This is often the case just because CNISs are so complex, and have so many different parts, each with their own special investigative techniques, evaluation measures, and so on. How to relate the results of evaluation of the different parts to one another is in itself a significant research question, which needs to be addressed if we are to be able to evaluate CNISs as whole systems.
- Research on how to incorporate users into the evaluation cycle (at all points, and with active participation of the users). Although action research and related methods have been used in a few instances of CNIS evaluation, by-and-large this is an unexplored area, which needs further to be investigated. This means going beyond such techniques as user-centered design, or participatory design, to enable (potential) users of CNISs to be significant evaluators throughout the design/evaluation cycle. How to do this is a significant research issue, which has yet to be resolved.

4th DELOS Workshop on DL Evaluation – Breakout Group on Metrics and Testbeds

Report of Breakout Group on Metrics and Testbeds

Chair of group:
Ingeborg T. Sølvberg

Members of group:
Norbert Fuhr, Sarantos Kapidakis, Christian Mönch, Noriko Kando (day 1), Ronald Larsen (day 1), Fillia Makedon (day 1), Zoltan Toth (day 1), András Micsik (day 2), Michalis Sfakakis (day 2).

The discussion was divided into two parts: Metrics and Testbeds.

Metrics

Setting the scene:
The participants agreed that a Digital Library has as main components: users, collections and system (technology); and usage, that depends on, as well as gives requirements for, all three components.

The discussion started with questions like and "do we need several levels for the metrics, and at least two: one for the DL itself and one for user interaction/satisfaction?", and "how can you know *what* you are measuring if you are observing users"?

The discussion centred very much about the user, as the opinion in the group was that the main objective for a digital library is to satisfy the user's needs.

But what does *user satisfaction* mean? Is it Satisfaction vs. Expectation? Is it possible to get the user's responses by automatic evaluation, such as logging the users activity and make the evaluations; or by observing the users using 'intelligent' cameras? If something can be automated, can it then also be verified?

The evaluation must be goal-driven and conducted in a systematic process. Four questions should be asked in the beginning of the evaluation process:

Who need this?
What shall be evaluated?
Why is this needed?
How can it be done?

The first steps in the process may be to get the users' goals and requirements, and to check if the goals are feasible. Then check if it is feasible to implement the goals and the requirements. If *yes*, check if and how the requirements are being influenced by the implementation.

Identifying the user group, that should be reached or served by the DL, is the first step in order to define the user requirements.

A Digital Library is developed for a specific use and for giving specific services.
The evaluation process of a DL starts by:

- Defining the goals for the Digital Library (Why and for whom is it made?)
- Defining the usage for the Digital Library
- Defining the success-criteria for how and when to say if the evaluation goals are fulfilled
- Defining the metrics for the success-criteria

And continues by

- Observing users/usage and collecting data

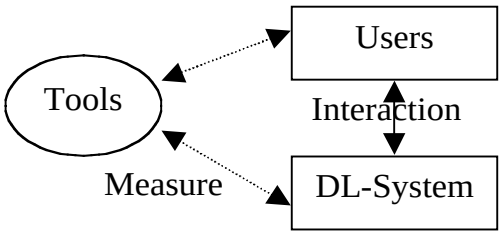
The process described above can give good evaluation results only if the Digital Library is used in the correct and planned way.

Specific methods and metrics have to be developed to measure user requirements, with focus on usage. A DL exists in changing environments, and this indicates that evaluation is a neverending process, and the goals (and metrics) may be reconsidered due to shifts in the usage of the DL.

We had a fairly abstract model of an *evaluation space*. The dimensions of this space are defined by the criteria, and the overall evaluation is the distance of the DL from the desired goal in the evaluation space.

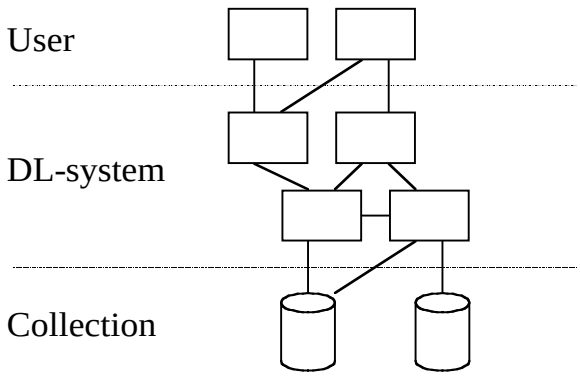
Testbeds

Due to the complexity of a DL-system (interaction of a large number of components, dependency on the nature of the content, dependency on the behavior of the users) a testebed has to be a complete DL-system that is equipped with measurement tools.



In order to test an individual component (or subsystem), all other components (or subsystems) have to be provided by the testbed. Individual components can then be integrated and compared. The basic types of components we identified (basic subsystems) were:

1. Collections of documents.
2. DL-system components
3. User components (generic, e.g. web-based, or DL-specific)



The group concluded that a Testbed consists of document collection(s) plus operations:

A Testbed is a digital library and an evaluation goal

In evaluation research using testbeds, the testbed should *not* be treated as a black box. In order to help to improve the overall system, each component should be available for inspection.
This requires that all components of a digital library are available: user, system, collection, and usage. One of the components can be selected for evaluation, and the component can be exchanged with another one for further research and testing.

A major problem is semantic interoperability between the components of the testbed; for example Systems and Collections. The collections may use proprietary semantics that deviates from other components.

Challenges with this approach are:

1. Generic components (collections and generic user components) need a well defined interface with a well defined semantics.
2. The decomposition should be subject to reconsideration because it may hinder the identification of new services.

Research needed. Summary.

There is a strong need for further research in this area. Some important research questions raised in the Breakout group are:

Research Questions concerning metrics

1. Which criterias besides user satisfaction are important?
2. How can user satisfaction be measured?

Research Questions concerning testbeds

1. Types of components (subsystems) have to be identified.
2. Interfaces between subsystems have to be well defined:
    a. Semantic aspects (interoperation, e.g. exchangability between different collections.)
    b. Syntactic aspects (Signatures)
3. Which kind of collections shall be provided

Report of the Breakout Session on Next-Generation Initiatives

Participants
Christine Borgman
Carol Peters
Noriko Kando
Javed Mostafa
Fillia Makedon
Ronald Larsen

Motivation and Background

The breakout session was charged to suggest directions for and attributes of a new initiative in evaluation of digital libraries. It began this exploration with a bit of reflection on past drivers for computational initiatives. The need for resource sharing and communication, for example, led to networking of computers, while the computational requirements underlying scientific challenges resulted in the development of supercomputing centers. More recently, the needs and opportunities for information access and organization on a global scale led to international initiatives in digital libraries. From the perspective of today's networked systems, we see a growing interest in distributed computation and information access reflected in concepts such as the computational grid. Difficulty in information usability on the Web is leading to new notions of semantically-aware services embodied in a "semantic web." The increasing globalization of business and defense requirements is driving the development of a new generation of cross-lingual information services.

Considerations Emerging from the EU 6th Framework

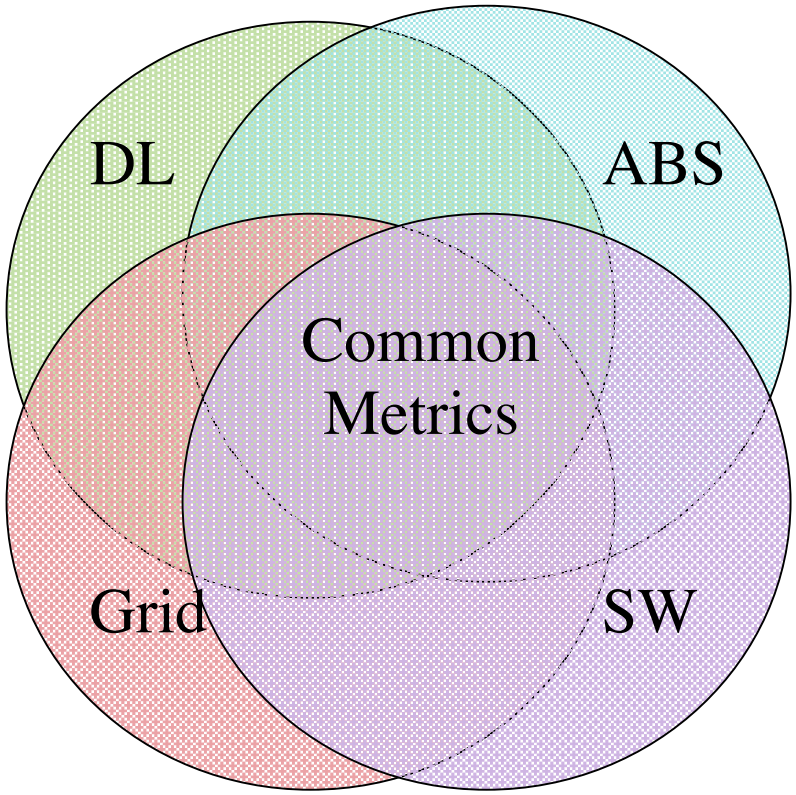
Lessons from the 5th Framework, now in its final stages, have included a growing appreciation for the difficulty of managing many small, distinct projects distributed among a sizable number of countries. The notion of an Information Society is attracting major funding in the 6th Framework, with a focus on integrated, large projects, operating in partnership with industry. A significant objective of the 6th Framework is to reduce the management overhead associated with these projects. Substantial clarity of thought is required, including:

- A clear understanding of the goal
- Elucidation of themes, directions, and expressions of interest from the greater community
- Appropriate measures and metrics to assess progress
- The ability to unambiguously measure progress
    - Time elapsed
    - Money spent
    - Performance achieved
    - Functionality developed
    - Users served

The development of the 6th Framework provides an opportunity for integrated thinking and understanding large-scale developments in an even broader context. For example, one could think of grid computing as distributed computation with information dependencies, while digital libraries manage distributed information with computational dependencies. Agent-based systems can be thought of as mobile entities seeking distributed information and computational services, while the semantic web can be thought of as striving for coherent coupling of distributed information with agent-based services.

Integrated Evaluation Strategies Proposed

The 6th Framework can capture a major opportunity to advance the techniques for evaluation of networked systems by taking an integrated approach that extends beyond digital libraries to include agent-based systems, grid computing, and the semantic web.



Each of these communities shares a mutual need for useful measures and faces similar difficulties in developing them. Instrumented test beds (either physical or virtual) are required, as are standardized metrics. A broad range of studies are required to understand how these systems actually perform in particular contexts, specific settings, and for diverse users.

Programmatic Drivers Identified

While it may be difficult to attract public resources and energy for measurement and evaluation of networked information systems such as digital libraries, there are very real societal motivations for so doing. These systems are increasingly becoming a foundation for education initiatives (consider, for example, the National Science Foundation's NSDL program). They are rapidly becoming the primary information infrastructure behind strategic defense interests (consider, for example, the heightened interest in Homeland Security and countering terrorism throughout the world). Networked information and computational systems are part of the critical infrastructure of the modern world, and much of the remaining critical infrastructure is fundamentally dependent on the underlying information infrastructure. Economic competitiveness and workforce productivity are also hindered without a comprehensive ability to drill down into information systems to understand how they are contributing to our society.

Common means for measuring effectiveness and efficiency are required that transcend any one information technology. We need rigorous and well-understood measures of usability and cost/benefit, and we have the opportunity to develop them in an integrated fashion, providing comparable measures across diverse technologies. But this justifiable interest in measurable results is too often either an unfunded mandate or an insufficiently funded afterthought. Rigorous efforts in information retrieval evaluation (the Tipster TREC workshops, for example) have clearly shown the value of these efforts as well as highlighting the labor-intensive nature of current approaches. They are time consuming and costly, but there is no natural law that says they have to be. Success requires an ability to attract appropriate talent and resources to the problems of evaluation, and developing the means of performing rigorous evaluation cost-effectively, and probably continually. Sufficient experience has been gained on narrowly focused evaluations that, if harvested and studied, can be used as leverage in the development of future test beds and evaluation tools.

Key Research Directions

The same issues and challenges from a performance perspective pervade the areas identified above (digital libraries, semantic web, grid computing, and agent-based systems). Issues of scale and scalability dominate; questions of interoperability among diverse and heterogeneous sources and systems challenge the best minds; usability of these systems by a broad range of individuals must be accommodated, entailing careful consideration of information content, computational tools, and user needs. A coherent set of evaluation standards and measures is needed.

Test beds provide a common reference point for evaluation studies, but these test beds must be well conceived and relevant to the problem domain. For the systems evolving today, distributed test beds of international scale and scope are required. They need to reflect the diversity of data encountered, services rendered, and the user groups served. A good test bed strategy will create a competitive tension and motivation to excel throughout the community. This has been accomplished successfully in the Text

Retrieval Conferences (TREC), which can provide a number of valuable lessons for building the next generation of test beds. A successful test bed project need not be a massive undertaking. I modest, relatively small project could be launched with the specific objective of engendering interest and creating momentum and a sense of progress. The open source model for software development could also provide an array of valuable lessons for building cooperative collaborations throughout a distributed and highly diverse world community.

International Collaboration

The European Union has been very successful in building consortial relationships across diverse populations. The European Commission has provided effective and coordinated leadership in these endeavors. In a very different model of collaboration, Asian cross-lingual information (Chinese / Japanese / Korean / English) projects have also been successful. The U.S. has built a substantial body of relevant experience through projects like TREC that have attracted international interest. The conclusion of this breakout session was that we all have much to learn and much to leverage from the collective experience of the world community. The richness of discussion and international understanding that follows from multinational development projects, alone, makes them worth the effort. But in the area of global information systems and services, such projects are a technological necessity.

June 12, 2002

Report on Breakout Group: Evaluating Digital Library Users and Interfaces

Group chaired by Christine Borgman

Members:

Jillian Griffiths
Laszlo Kovacs
Javed Mostapha
Michalis Sfakakis
Linda Banwell

Scoping:

Users and interfaces encompasses user behavior, uses of the DL, conceptual and technical asepcts of user interfaces.

Interfaces is a term used in two distinct ways in DLs: (1) in the human-computer interaction sense, in which the user interface is the manifestation of the DL functions and services and thus is deeply embedded in the system; and (2) in the software engineering sense, as the boundary between system components, such as the interface between client and server. Although we are largely focused on the first definition in evaluation of digital libraries, both definitions are relevant considerations.

Research questions:

A few basic questions indicate the complexity of this area and the issues to be addressed in evaluation:

Who are the users of the DL?

What are the uses of the DL?

What is the relationship between the users and the uses?

What are the specific goals of the DL?

How can we apply methods and measures frm related areas such as HCI, IR, information seeking, information needs and uses to the evaluation of DLs?

How to evaluate DLs that cross domains, languages, and cultures?

How to adapt evaluation methods to different formats (e.g., text, audio, image, video) and mixed content?

How to evaluate multi-modal environments (e.g., textual, visual, audio, gestures, etc.)?

How to determine the appropriate level of granularity for evaluation? Evaluation may be context and goal dependent.

Research methods:

Different methods will be required for different evaluation goals: economic, productivity, learnability, social impact, domain specific goals (e.g. scientific thinking, information retrieval).

DL evaluation should be holistic, and recognize that DLS have a larger scope that IR, and for example may include creating and using information resources, as well as searching for them.

Qualitative and quantitative methods are required, and will have different reliability and validity criteria.

Criteria for determining the best research questions and methods for evaluating User and Interface aspects of DLS:

Cost of evaluation

Cost benefit of evaluation (e.g., how much money is saved through productivity improvements as a ratio of the development cost of the system?)

Methods must be flexible and adaptable to new user interface paradigms, new models fo uses and users

Ability to share methods, instruments, and testbeds.

Validity and reliability (e.g. tradeoffs between evaluation in situ and evaluation in laboratory settings).

SUMMARY

The Breakout group on Evaluating Digital Library Users and Interfaces focused on user behavior, uses of DLs, and conceptual and technical aspects of user interfaces. This is a complex area, as not only do users and uses of DLs need to be distinguished, but many other variables must be considered, such as domains, languages, and cultures; multiple formats of content such as text, audio, image, and video; multi-modal DL environments such as wired and wireless, visual, audio, gesture-based interfaces; and the granularity of evaluation. Different methods will be required for different evaluation goals, such as economic, productivity, learnability, social impact, domain specific goals (e.g. scientific thinking, information retrieval). We focused on the criteria for determining appropriate evaluation of DLs, such as cost of evaluation; cost benefit of evaluation (e.g., how much money is saved through productivity improvements as a ratio of the development cost of the system?); the ability to share methods, instruments, and testbeds; and validity and reliability (e.g. tradeoffs between evaluation in situ and evaluation in laboratory settings).

APPENDIX

APPENDIX - A

Workshop Program

Thursday, 6 June

9-9:15	Workshop opening: László Kovács, Ingeborg Sølvberg, Christine Borgman
9:15-10:45	**Setting the Background on DL Evaluation: Reports from working groups** (Chair: Ingeborg T. Sølvberg)
9:15-9:35	LONG: Ronald L. Larsen. The DLib Test Suite and Metrics Working Group: Harvesting the Experience from the Digital Library Initiative.
9:35-9:55	LONG: Michael Mabe. DL Classification & Evaluation: A Publisher's View of the Work of the DELOS Evaluation Forum.
9:55-10:15	LONG: Noriko Kando. Evaluation of Information Access Technologies.
10:15-10:45	Discussion
10:45-11:15	Coffee
11:15-12:15	**Users and User Interfaces** (Chair: Preben Hansen)
11:15-11:35	LONG: Nicholas J. Belkin. A Framework for Criteria and Measures for Evaluation of User Interfaces in Digital Libraries.
11:35-11:45	SHORT: James Ford et.al. Evaluation Metrics for User-Centered Ranking of Content in MetaDLs.
11:45-11:55	SHORT: Michalis Sfakakis and Sarantos Kapidakis. Evaluating User Behavior on Data Collections in a Digital Library.
11:55-12:15	Discussion
12:15-1:30	Lunch break
1:30-3:15	**Evaluation in Context** (Chair: András Micsik)
1:30-1:50	LONG: Fabio Abbattista et.al. Virtual Agents for a Bookstore: an Empirical Evaluation.
1:50-2:10	LONG: Ann Bishop and Bertram Bruce. Digital Library Evaluation as Participative Inquiry.
2:10-2:30	LONG: Christine L. Borgman. Evaluating a Digital Library for Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype (ADEPT).
2:30-2:50	SHORT: Evans, O'Dwyer, Schneider. Usability Evaluation in the Context of Digital Video Archives.
2:50-3:15	Discussion
3:15-3:45	Coffee
3:45-5	**breakout groups**
5-5:45	**Reports from breakout groups**

Friday, 7 June

9-10:45	**Metrics and Testbeds** (Chair: Javed Mostafa)
9-9:20	LONG: Carol Peters. Creating a Multilingual Test-Bed for Cross-Language System Evaluation.
9:20-9:50	LONG. Norbert Fuhr. Evaluating Efficiency vs. Effectiveness for Vague Queries and Similarity Search in Digital Libraries.
9:50-10	SHORT: Robert J. Sandusky. Digital Library attributes: Framing research and results.
10-10:10	SHORT: Christian Mönch. On the Assessment of Scalability of Digital Libraries.
10:10-10:45	Discussion

10:45-11:15	Coffee

11:15-12	**Evaluation of DL services and scalability** (Chair: Christine Borgman)
11:15-11:35	LONG: Jillian R. Griffiths and Shelagh Fisher. A Quality Management Approach to the Evaluation of Digital Library Services.
11:35-11:45	SHORT: Linda Banwell. TEL: The European Library - The Gate to Europe's knowledge.
11:45-12	Discussion

12-1	**breakout groups**

1-2:15	Lunch

2:15-3:30	**Reports from breakout groups, conference wrap-up, next steps**

3:30	Coffee

APPENDIX - B

List of Participants

Name	Company
Bánhegyi, Zsolt	Library of the Hungarian Academy of Sciences, Hungary
Banwell, Linda	Information Management Research Institute, School of Information Studies, Northumbria University, England
Belkin, Nicholas J.	Rutgers University, New Brunswick, USA
Bishop, Ann Peterson	University of Illinois, Champaign, USA
Borgman, Christine	University of California, Los Angeles, USA
Dobó, Katalin	Central European University / Open Society Archives, Hungary
Fuhr, Norbert	University of Dortmund, Germany
Glushakov, Sergey	OSA, Open Society Archives at Central European University, Hungary
Griffiths, Jillian R.	CERLIM, Manchester Metropolitan University, UK
Hansen, Preben	SICS, Sweden
Kando, Noriko	National Institute of Informatics, Tokyo, Japan
Kapidakis, Sarantos	Ionian University, Athens, Greece
Kovács, László	MTA SZTAKI, Hungary
Larsen, Ronald L.	University of Maryland / OIT / MAITI, USA
Lops, Pasquale	Department of Computer Science - University of Bari, Italy
Mabe, Michael	Elsevier Science & City University, Oxford, UK
Makedon, Fillia	Dartmouth College, Hanover, USA
Micsik, András	MTA SZTAKI, Hungary
Mostafa, Javed	Indiana University, Bloomington, USA
Mönch, Christian	Computer Science Dept. - NTNU, Trondheim, Norway
Peters, Carol	IEI-CNR, Pisa, Italy
Saari, Seppo	Finnish Higher Education Evaluation Council, Helsinki, Finland
Sandusky, Robert J.	Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Chicago, USA
Schneider, Stephan	Tecmath AG, Kaiserslautern, Germany
Sfakakis, Michalis	National Documentation Centre – NDC, Athens, Greece
Sølvberg, Ingeborg T.	The Norwegian University of Science and Technology, Department of Computer and Information Science, Trondheim, Norway
Tapolcai, Ágnes	National Szechenyi Library, Hungary
Tóth, Zoltán	MTA SZTAKI, Hungary