

# Digital Libraries: A Generic Classification and Evaluation Scheme

Norbert Fuhr<sup>\*</sup>, Preben Hansen<sup>\*\*</sup>, Michael Mabe<sup>\*\*\*</sup>, Andras Micsik<sup>†</sup>,  
Ingeborg Sølvsberg<sup>‡</sup>

**Abstract.** Evaluation of digital libraries (DLs) is essential for further development in this area. Whereas previous approaches were restricted to certain facets of the problem, we argue that evaluation of DLs should be based on a broad view of the subject area. For this purpose, we develop a new description scheme using four major dimensions: data/collection, system/technology, users, and usage. For each of these dimensions, we describe the major attributes. Using this scheme, existing DL test beds can be characterised. For this purpose, we have performed a survey by means of a questionnaire, which is now continued by setting up a DL meta-library.

## 1 Introduction and background

“What is a digital library?” The answer depends upon whom you are asking. This is even truer if you ask, “What is a *good* digital library”. Several research disciplines (e.g. library science, computer science, sociology) and groups of practitioners (e.g. publishers, librarians) are interested in and make contributions to digital libraries (DLs). Each of them has a different view on DLs, and is focusing on those aspects that are relevant from that specific viewpoint.

### 1.1 Digital libraries and evaluations

[1] identifies two “schools” which have different views and approaches towards DLs; the research community and the (traditional) library community.

The research community focuses upon the information content that are collected and organised in order to fulfil the needs of (selected) user groups; i.e. digital objects, architecture, and usage.

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 on Digital Libraries”, 1997]

---

<sup>\*</sup> University of Dortmund, Germany. Email: fuhr@cs.uni-dortmund.de

<sup>\*\*</sup> Swedish Institute of Computer Science, Kista, Sweden. Email: preben@sics.se

<sup>\*\*\*</sup> Elsevier Science Ltd., Oxford, UK. Email: m.mabe@elsevier.co.uk

<sup>†</sup> MTA SZTAKI, Budapest, Hungary. Email: micsik@sztaki.hu

<sup>‡</sup> Norwegian University of Science and Technology, Trondheim, Norway. Email: Ingeborg.Solvberg@idi.ntnu.no

The library community looks upon DLs as institutions or organisations that offer information services in digital form, and how existing structures can adapt to new technology and new challenges.

DLs are organisations that provide the resources, including the specialised staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. [Digital Library Federation (DLF), 1998]

The definition of a DL varies, and that is reflected in the questions on evaluation of DLs:

- What can be evaluated?

For example, a librarian may focus on the collection, whereas a computer scientist may be interested in the technological aspects only, irrespective of the content of a DL. On the other hand, an institution wanting to subscribe to a DL may want to choose the best among several DLs with similar content, thus taking a user-oriented view on a DL.

- What and how to measure?

Some system designers may focus on the efficiency of a DL system (i.e. usage of computational resources), whereas others are interested in effectiveness. For the latter, one could e.g. consider relevance only and apply the standard information retrieval measures of precision and recall; a broader view would look at typical tasks to be solved with the DL system and measure e.g. task time and completion rate.

- Who needs the results from the evaluation?

In many cases, results are needed for decision-making: For example, the management of a library has to select a new DL software; a librarian managing subscriptions looks for DLs offering the content mostly needed by his clients; a system developer has to make design choices.

- When is it appropriate to evaluate?

Evaluations may take place at any place in time: A system developer may have to choose between several methods for performing a certain function, for which a rather focused evaluation in a laboratory setting may suffice. Decisions for selecting a piece of software or subscribing to a DL should be based on the final product.

As stated in [13], any evaluation has to meet certain requirements. Hence, a set of important issues has to be considered, such as the construct for evaluation, the context of evaluation, the criteria, the measures and the methodology.

However, Saracevic and Covi [13] conclude that there are no clear agreements regarding the elements of criteria, measures, methodologies as well as of the larger “view” which involve the construct and context of evaluation. We therefore see our paper as a step in the direction of resolving some of these issues.

The US DLI test suite<sup>1</sup> is a group of six digital library test beds developed by the projects of the first phase of the US Digital Library Initiative. These

<sup>1</sup> <http://www.dlib.org/test-suite/index.html>

test beds comprise different media; however, since the focus of most the projects spawning off the test beds was on technological aspects, users and usage as well as the content play a minor role in most of these test beds. Related to this effort, the D-Lib Working Group on Digital Library Metrics<sup>2</sup> was formed and was involved in the organisation of a workshop<sup>3</sup> in 1998, which addressed several aspects of DL evaluation. Unfortunately, this effort has not been continued.

The Digital Library Evaluation Forum of the DELOS Network of Excellence<sup>4</sup> aims at providing an infrastructure for the evaluation of performance related aspects in accessing digital libraries. Research into DLs needs large test beds to evaluate and demonstrate new concepts. In recent and coming years, several excellent collections have been/will be created with EU funding, and the collections will be integrated in new applications. For defining or describing a DL test suite, the dimensions of such a test bed have to be defined.

In this paper written by several members of the DELOS working group<sup>5</sup> “Digital Library Test Suite”, we first give a brief survey on related work (section 2). Then we outline the general idea of our holistic approach to DL evaluation (section 3), followed by a more detailed presentation of our description scheme (section 4). Section 5 describes the results of the evaluation of the scheme and an ongoing effort for building a DL meta-library. Finally, we give an outlook on future work in this area.

## 2 Related work

Methods and tools to evaluate computer systems have been investigated for several decades, with a special focus on basic parameters for measuring the performance of a system: effectiveness and efficiency. The cost factors can be calculated indirectly.

In the area of information retrieval (IR), evaluation plays a central role since many years. The TREC (Text retrieval Conference) initiative is an ongoing effort for developing standard benchmarks for IR methods and systems. In the different tracks of TREC [14], a variety of collections (Web documents, newspaper and newswire articles, spoken broadcast news) and uses (ad-hoc queries, interactive querying, filtering, question answering) are considered. As evaluation criterion, mainly retrieval effectiveness (in terms of recall and precision) is regarded. From a DL point of view, the collections employed in TREC lack the rich structure and inter-document relationships that are typical for DL collections. Furthermore, involvement of real end-users is only marginal (in the interactive track).

Common approaches within the HCI research area are different usability evaluation methods ([10], [7]). Furthermore, in the process of assessment of users and their interaction with computers, HCI have used a diverse set of methods and techniques. These methods range from controlled laboratory-based settings,

<sup>2</sup> <http://www.dlib.org/metrics/public/index.html>

<sup>3</sup> <http://www.dlib.org/metrics/public/6-98-workshop/index.html>

<sup>4</sup> <http://www.ercim.org/DELOS>

<sup>5</sup> [http://www.sztaki.hu/delos\\_wg21](http://www.sztaki.hu/delos_wg21)

simulated situations with simulated tasks to longitudinal workplace studies (e.g. [12]). From a HCI point of view, techniques such as usability inspection methods, cognitive walk-through, task analysis methods, think aloud, ethnographic methods etc may be used in order to evaluate the use and usage of DLs. Usability studies (ease of operation) can be distinguished from usefulness (serving an intended purpose) even if the two are hard to separate in the context of evaluation [8].

Following these ideas, [9] describes a human-centred approach for designing DL systems. Here the focus is on assessing human information needs and corresponding tasks, thus evaluation deals with the effects of the DL on the subsequent human information behaviour. As a consequence, the authors claim that DL design should be process-oriented and iterative rather than product-oriented and summative.

A digital library has many similarities with a traditional, physical library, but it has also many differences. [2] distinguishes evaluation procedures in a physical library from those in a digital library. She reviews usability evaluation work within taxonomy of system design, development and deployment. In addition to an evaluation strategy where evaluation data are collected throughout the system life cycle, she includes a second strategy where the evaluation methods themselves are evaluated. This is the evaluation approach in the Alexandria Digital Library (ADL) [5]. In this study the user reactions to the ADL interface and to the functionality and content of ADL was evaluated, as well as the user characteristics used and the study approach itself.

With the current changes in the library area, the problems of library statistics and performance measurement have gained increasing attention<sup>6</sup>. [4] distinguishes five dimensions of evaluation: extensiveness, effectiveness, efficiency, costing and quality. [3] claims that the traditional notions of efficiency and effectiveness should be balanced with the benefit to customers ('value'), and describes a conceptual framework for measuring the latter. [11] presents lists of DL evaluation metrics (quantitative data) for two facets, namely users and data/services. In addition, so-called nuggets serve for collecting qualitative data relating to a variety of topics.

The different evaluation needs described above imply that both qualitative and quantitative data should be collected and analysed in order to contribute to our understanding of how DLs should to be designed in order to support different users and usages. Generally, we may say that from the level of system performance evaluation to the evaluation level of the situational and contextual factors the different methods and techniques range from quantitative data collection methods (e.g. TREC-based performance measures) to qualitative data collection methods such as ethnographic AI observations.

This calls for building a collection of evaluation methods and techniques that may be used within the framework of DL evaluation. Different evaluation methods may be used for different purposes and levels of analysis.

---

<sup>6</sup> See e.g. the bibliography at <http://web.syr.edu/~jryan/infopro/statgen.html>.

### 3 A holistic approach to DL evaluation

A DL is a special kind of an information system, and consists of several components such as a collection, a computer system (a technical system), persons, and the environment (or usage), for which the system is built. For DLs it is important to integrate systems, information/collection and humans, as well as support the viewpoints of the different shareholders in DL research. Thus, a librarian may focus on the collection, whereas a computer scientist may be interested in the technological aspects only, irrespective of the content of a DL. However, ignoring the other dimensions of the problem may lead to impractical solutions. For example, in the area of information retrieval, most technological-oriented research assumes a batch-like environment, whereas the few evaluations targeted towards interactive retrieval have not yet been able to show that technological differences really matter [14].

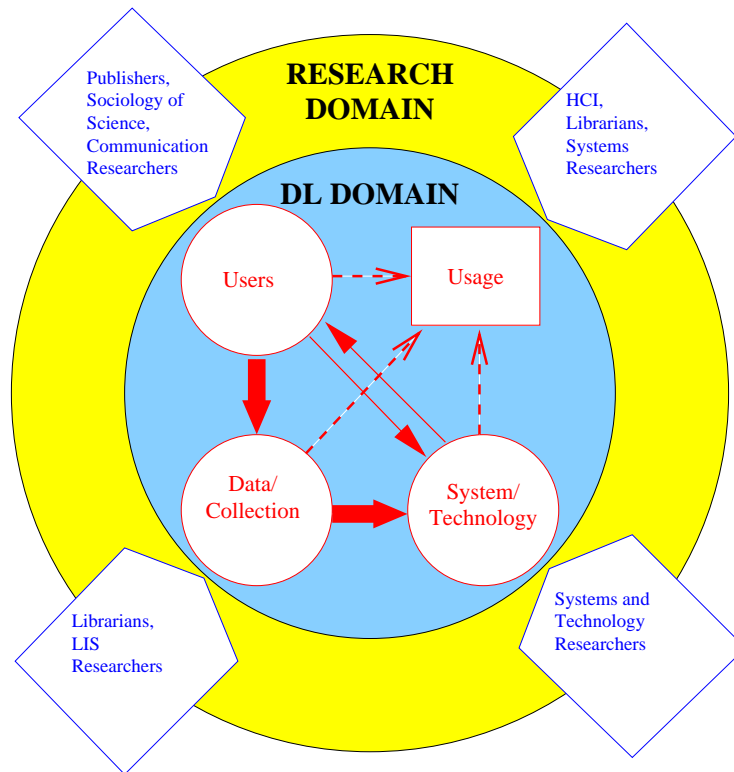
Important issues to be considered from a broader viewpoint of DL evaluation would include the following:

- The underlying system and its components (this involves e.g. classical information retrieval evaluation methods and techniques as well as overall systems performance).
- The interface and interaction level of the activities between the user and the system (this involves classical human-computer usability evaluation issues).
- Support for different access and usage strategies (e.g. analytical search, browsing, navigation, bibliographic search, collaboration, annotations).
- The work tasks should be supported. Often, only the task of searching is supported in the design of an access system.
- Situational and contextual factors of DLs are important, such as organisational and group issues.

When comparing these requirements with the outcomes of previous DL programmes, it becomes apparent that there is an imbalance in satisfying the different DL research domains, with a predominance of technology related initiatives.

In order to take a broader view on DL evaluation, our working group decided to focus on the development of test suites that satisfy the needs of all kinds of DL researchers. As a result of a brainstorming exercise, we developed the diagram shown in figure 1.

Our new approach uses a generic definition of a digital library as its starting point, and is illustrated by the small circles within the central circle labelled “DL DOMAIN”. The model falls into three non-orthogonal components: the users, the data/collection, the technology used. Definition of the set of users predefines the appropriate range and content of the collection (block arrow connecting “users” to “collection”). The nature of the collection predefines the range of appropriate technologies that can be used (block arrow from “collection” to “technology”). Finally, the attractiveness of the collection to the users and, secondarily, the ease of use of the technologies by the user group, determine the extent of the usage of the digital library (thin arrows show the human-computer interactions, while the dotted arrows show the collective contribution of user, collection and technology interactions to observed overall usage).



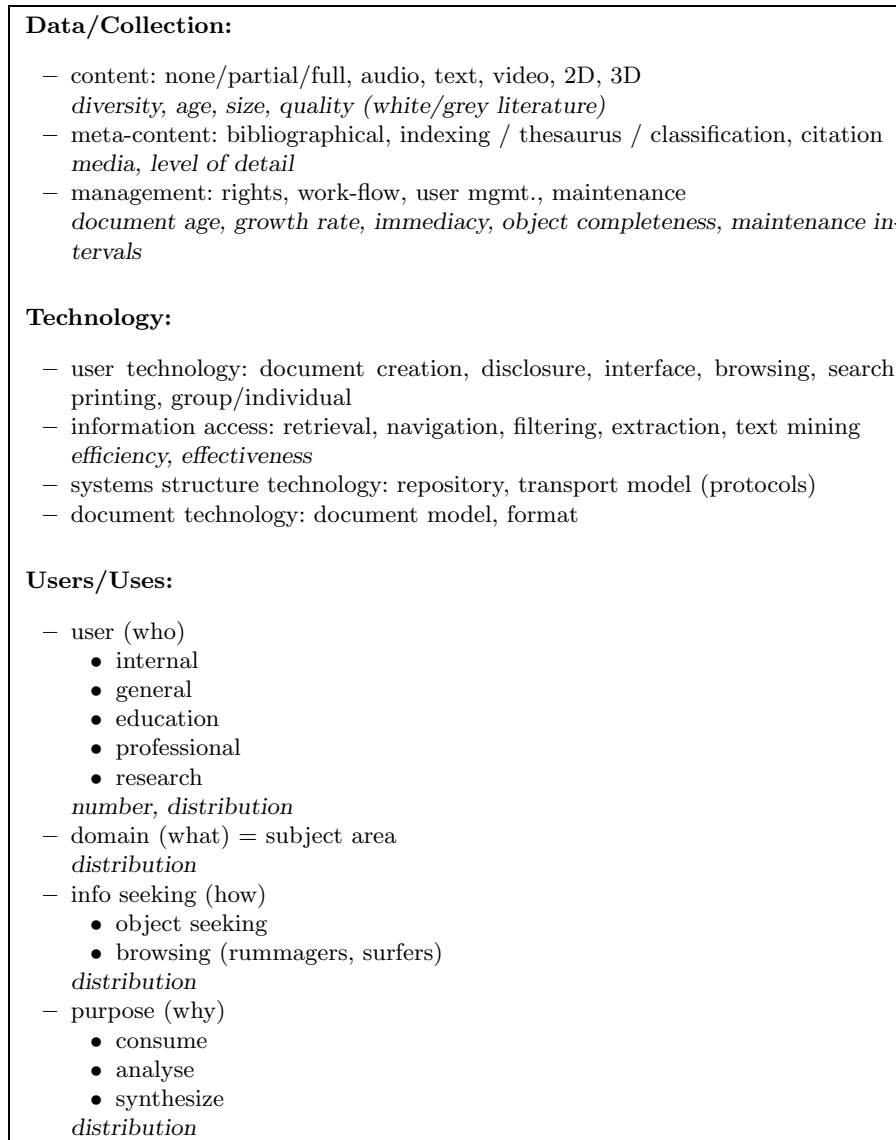
**Fig. 1.** A Generalised Schema for a Digital Library

From this starting point, it is possible to move outwards to the domain of the DL researchers (outer ring), and to use the non-orthogonal relationships between the principle research areas (users, usage, collection and technology) to create a set of researcher requirements for a DL test bed. Because “content is king”, the nature, extent and form of the collection predetermine both the range of potential users and the required technology set.

## 4 Evaluation Criteria and Metrics

### 4.1 Categories of description scheme for DLs

While many of the descriptive characteristics of DLs are interconnected and dependent (the nature of the user predefines the collection they will use, which in turn delimits the potential tool set they may adopt), it is possible to assign relatively independent criteria within the overall domain of data/collection, technology and users. Creating such a definition allows us then to select denumerable subsets as appropriate evaluation criteria.



**Fig. 2.** Description scheme and evaluation criteria for digital libraries

As a refinement of the concepts developed in the previous section, we identified the major parameters characterising the three dimensions of the DL domain. These parameters are listed in figure 2. Most of these parameters are either binary (yes/no) valued or have a restricted set of values. A few additional parameters have numeric values, which are shown italics here. Below, we describe the parameters for the different dimensions in more detail.

## 4.2 Users and uses

The uses and the user are intimately connected with the four basic questions that can be asked about any market. Who is the market? What are they interested in? How and why do they behave as they do?

The “who?” question is largely a matter of demographics and hierarchy within the information chain. In the first cut, users are either internal to the DL system or external. In the case of the external users, they correspond to different levels of the standard information pyramid: the mass market (general), primary, secondary and tertiary educational market, the industrial, manufacturing or professional users, the high level university, corporate or institutional research users. Following this classification of user demography we can evaluate in terms of the numbers of each user type and their distribution among the user classes.

The “what?” question concerns the subject area of interest to a user. That is the domain of their DL use. For evaluation we can use the distribution of subject areas as a metric.

The third dimension relates to the ways in which users seek information, the “how?” question. Users can adopt essentially two strategies. The first is direct object-seeking, that is the use of sophisticated tools (largely search engines) to identify specific, singular pieces of information that resolve closely defined questions. The second is the traditional wandering approach of library browsing. There may or may not be more systematic approaches contained within this strategy. A user may use a classification scheme or other labels to limit the domain of the browse. Alternatively they may randomly wander around the information lighting upon topics of interest serendipitously. For evaluation purposes, the distribution of users between these approaches can be used.

Lastly, we can consider the purpose behind a users information encounter, the answer to the “why?” question. For some users the encounter may simply be to consume the information for pleasure or interest. For others the information may be an object to analyse critically for educational, research or review purposes. For another group, the information will be crucial to synthesize new works via quotation, commentary, annotation or citation. Again for evaluation purposes, the distribution of uses between these categories can be a useful metric.

## 4.3 Data/Collection

The collections and the information objects in a Digital Library can be described using different axes: content description, quality/reliability qualifiers, and management and accessibility qualifiers.

The collections in a digital library contains information objects gathered according to some rules or ideas, on the basis of one or several attributes to be described collectively. It may be a thematic collection such as work by a specific author or composer (or “creator”) (W. Shakespeare, J.S. Bach); or subject (mathematics, history); a collection based upon media types (paper, CDs, films, maps, a.o.); age (information objects ‘produced after 1968’), or just a general

collection for a wide audience where the collection may include a variety of media types. To describe quality in an objective manner is almost impossible. It is, however, feasible to give descriptors that may help to estimate quality and authenticity. In scientific domains it may be of importance to know if a collection contains 'grey' or reviewed literature, and if the collection's owner is well reputed, by giving the name of the owner.

The collections may contain primary objects like the text of Shakespeare's "Romeo and Juliet", or the film "Cinderella". Collections of secondary objects contain bibliographic descriptions; holdings; data to assist in authority control (thesauri, gazetteers, classification schemes, etc.), or may assist in the thematic information seeking process (collections of citations). The metadata scheme(s) used to describe the information objects gives the level of detail of the data (MARC format, Dublin Core, RFC1807, robot generated, none).

A collection needs maintenance. Redundant information objects have to be removed, errors have to be repaired, and the growth of a collection must be secured. A responsible organisation or body must be in charge of this work. Additional functions that need to be handled properly are user management, security and access control. Examples of possible qualifiers may be the name(s) of the body in charge; maintenance intervals; statistics of growth rate, accessibility, number of users, types of users, and others.

#### 4.4 Technology

The technological issues can be subdivided into four areas, namely user technology, information access, systems structure and document technology.

User technology deals with the functions that the DL system offers to the user: Most basic, these functions have to be provided via an appropriate user interface. Documents are made accessible via searching and browsing; furthermore, there may be a disclosure mechanism that notifies the user about new documents that might be relevant for him. Once a relevant document is located, most users prefer to read it on paper; thus, a printing function is essential. Since users often work in teams, support for user groups also is an important function in DL systems, e.g. for collaborative filtering. Besides accessing existing documents, a DL system also may support the creation of new documents.

For information access, a DL system should implement a rich set of functions. Retrieval searches for documents in response to a query. Navigation follows (explicit or implicit) links between documents and/or metadata. Based on a profile specified by the user, filtering locates potentially relevant documents in a stream of incoming documents. Information extraction generates facts from text documents. Based on this input, text mining can discover correlations and trends in a document collection.

Systems structure technology deals with the architecture of the repository (e.g. centralised/distributed database, relational/object-oriented database management system, middle-ware) and the transport model (protocols for communication between the system and the user interface or between system components).

Document technology addresses the issue of the representation of documents. The document model describes the abstract structure of documents such as the hierarchical/hyperlinked logical structure, content media, layout, semantic content and external attributes. The document format specifies the syntax of the internal document representation (e.g. postscript, PDF, RTF).

## 5 Evaluation of the scheme: questionnaires and lessons learned

The scheme described previously can serve as a basis of various evaluation and classification efforts. By attaching answers to the questions in the scheme, a specialised description of a digital library can be received. These descriptions can be used to compare digital libraries or to select some for a specific testing purpose. As a first trial of these ideas, a survey was done in the second half of 2000 with a questionnaire reflecting main ideas of the classification scheme.

### 5.1 First survey

In order to gather a first set of information about Digital library collections and test collections, we designed a two-part questionnaire. The Questionnaire A (available digital libraries and test collections) concerned with the availability of existing digital libraries and test collections that could be used for research in the field of digital libraries. The Questionnaire B (desired digital library test collections) investigated future requirements of digital library test collections.

Questions in both questionnaires were put in a similar way, the wording was only changed to reflect the different target: existing DLs versus requirements and research needs. A set of questions corresponded to each of the three main categories of the scheme: Users/Uses, Data/Collection and Technology. Additionally, the gender and the work domain were asked about the person who responded. Answering questions was possible by selecting a single or multiple choices from a choice list and by optionally giving a comment. Naturally, respondents could abstain from answering some of the questions. Questionnaire A contained 31 questions and Questionnaire B contained 21 questions. (The reason why Questionnaire B had less questions is that some questions had no relevance to put for a not yet existing digital library.)

The survey was carried out by the open, Web-based survey tool of SZTAKI [6]. The survey was announced at various mailing lists with major audience of digital library researchers and developers, and our estimation is that roughly 3-4% of the targeted audience responded. Nearly 70% of the respondents were from the research domain, and the users of the evaluated digital libraries also had more than double weight in the research domain than in any other domain.

These surveys showed that while the proposed classification scheme seems to be appropriate for DL characterisation, the wording of the questions is rather problematic. Due to the holistic approach of the classification scheme, many different research areas are covered, and these areas have their different term

sets and language usage, which makes it difficult to create questionnaires which are equally understandable by researchers of different areas. Naturally, creating questions about this lively and multidisciplinary research area is not an easy task. We received several suggestions either via comments in the survey or via e-mail, which showed that an additional problem is to guess a reasonable granularity of choices for questions, which on the other hand does not hide new or unorthodox approaches, as these have great importance in a research survey. The low response rate showed that the return of the investment was not clearly seen by research communities. The possibility that filling these surveys could be the starting point of research cooperations was not emphasised by the survey environment on the required level.

## 5.2 MetaLibrary

The DL MetaLibrary<sup>7</sup> is the current effort of the DELOS WG2.1 working group, where lessons learned from the survey and new ideas are taken into account. This is an extensible survey database, where each DL collection/test-bed can register and provide information about itself. Information provided here can be updated any time by the original submitters after a password-based authentication. Questions from the old survey were redesigned and more opportunity was given for answers in free text. Questions are mapped on a hierarchy suggested by the proposed classification scheme. Thus, nodes in this hierarchy represent a research area or a DL functionality. These are called slots and they are identified by a unique number sequence. Questions and answers are stored in their corresponding slots, thus new questions can be easily introduced into the existing survey database. Slots can also be used for enhanced searching and browsing of the contents of MetaLibrary. Users will be able to browse the hierarchy of slots, get summaries or statistics of available solutions in selected slots, or search for DLs having a certain solution in a selected slot. With a sufficient coverage of the activity in the area, MetaLibrary can help DL people to find systems, test-beds and research partners for their needs, and at the same time lacking features or holes in research can also be identified.

## 6 Conclusions and outlook

In this paper, we have argued that evaluation of DLs should be based on a broad view of the subject area. For this purpose, we have developed a new description scheme using four major dimensions (collection, technology, users, usage) and we have described the major properties along each of these dimensions. Using this scheme, existing DL test beds can be characterised. For this purpose, we have performed a survey by means of a questionnaire, which is now continued by setting up a DL meta-library.

In order to ease the maintenance of this meta-library, we will investigate the possibility of switching to a harvesting scheme, where each provider of a DL test

---

<sup>7</sup> [http://www.sztaki.hu/delos\\_wg21/metelib/](http://www.sztaki.hu/delos_wg21/metelib/)

bed maintains a metadata record describing his/her test bed, and the harvester collects these records on a regular basis

Once we have metadata about a substantial number of existing test beds, we will relaunch our questionnaire on desired test collections: Given the existing test beds and the current interests of DL researchers, what types of test beds are still missing? The outcome of this poll may lead to an effort for building appropriate test collections, similar to the TREC and CLEF initiatives.

**Acknowledgement** This work was supported in part by the Network of Excellence DELOS funded by the European Commission.

## References

1. C. Borgman. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. MIT Press, 2000.
2. B. Buttenfield. User evaluation for the alexandria digital library project. Allerton Inst. WS, <http://edfu.lis.uiuc.edu/allerton/95/s2/buttenfield>, 1995.
3. J. Cram. Six impossible things before breakfast: A multidimensional approach to measuring the value of libraries. In *Proc. 3rd Northumbria Intl. Conf. on Performance Measurement in Libraries and Information Services*, pages 19–29, Newcastle upon Tyne, 1999. Information North.
4. P. Herson and E. Altman. *Service quality in academic libraries*. Ablex, Norwood, NJ, 1996.
5. L Hill, Ron Dolin, James Frew, R.B. Kemp, M. Larsgaard, D.R. Montello, Mary-Anna Rae, and J. Simpson. User evaluation: Summary of the methodologies and results for the alexandria digital library. In *Proc. ASIS*, pages 225–243, Medford, NJ, 1997. Information Today .
6. L. Kovacs and A. Micsik. A public service for surveys and decision polls. In *Proc. DEXA 2000*, pages 307–311, September 2000.
7. Roberta Lamb. Using online information resources: Reaching for the \*.\*’s. In *Digital Libraries ’95 Conference Proceedings.*, New York, 1995. ACM.
8. T. K. Landauer. *The Trouble with Computers: Usefulness, Usability, and Productivity*. MIT Press, Cambridge, Mass., 1995.
9. G. Marchionini, C Plaisant, and A. Komlodi. The people in digital libraries: Multifaceted approaches to assessing needs and impact. In *Digital library use: Social practice in design and evaluation*. MIT Press, Cambridge, Mass., 2001. (in press).
10. J. Nielsen and R.L. Mack, editors. *Usability Inspection Methods*. John Wiley & Sons, New York, 1994.
11. Catherine Plaisant and Anita Komlodi. Evaluation challenges for a federation of heterogeneous information providers: The case of nasa’s earth science information partnerships. In *Proc. WET ICE 2000, IEEE WS on Evaluating Collaborative Enterprises*, 2000.
12. J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey. *Human-Computer Interaction*. Addison Wesley, Reading, Mass., 1994.
13. T. Saracevic and L. Covi. Challenges for digital library evaluation. In *Proceedings ASIS*, volume 37, pages 341–350, 2000.
14. E. Voorhees and D. Harman. Overview of the eighth text retrieval conference (trec-8). In *The Eighth Text REtrieval Conference (TREC-8)*, pages 1–24. NIST, Gaithersburg, MD, USA, 2000.