Int J Digit Libr (2007) 8:21–38
DOI 10.1007/s00799-007-0011-z

REGULAR PAPER

Evaluation of digital libraries

Norbert Fuhr · Giannis Tsakonas · Trond Aalberg · Maristella Agosti · Preben Hansen · Sarantos Kapidakis · Claus-Peter Klas · László Kovács · Monica Landoni · András Micsik · Christos Papatheodorou · Carol Peters · Ingeborg Sølvberg

Published online: 15 February 2007

Abstract Digital libraries (DLs) are new and innovative information systems, under constant development and change, and therefore evaluation is of critical importance to ensure not only their correct evolution but also their acceptance by the user and application communities. The Evaluation activity of the DELOS Network of Excellence has performed a large-scale survey of current DL evaluation activities. This study has resulted in a description of the state of the art in the field, which is presented in this paper. The paper also proposes a new framework for the evaluation of DLs, as well as for recording, describing and analyzing the related research field. The framework includes a methodology for the classification of current evaluation procedures. The objective is to provide a set of flexible and adaptable guidelines for DL evaluation.

1 Introduction

Digital libraries (DLs) are complex systems; they can be, and are, viewed from different perspectives. The methods and metrics for the evaluation of DLs may vary according to whether they are viewed as institutions, as information systems, as new technologies, as collections, or as new services.

A DL is a particular kind of information system and consists of a set of components, typically a collection (or collections), a computer system offering diverse services on the collection (a technical infrastructure), people, and the environment (or usage), for which the system is built. When designing a DL the starting points are its intended usage and the corresponding user needs. The model is based upon the assumption that the user and the user needs specify the main requirements with respect to the range and content of the collections. The nature of the collections will thus predetermine the range of technologies that are needed. The attractiveness of the collections to the users and the ease of use of the technologies by the user group will determine the extent of the usage of the DL.

DLs are components in several types of applications in areas such as cultural heritage, health, government, learning, and science. Technological advances in areas like information searching and retrieval, information storage, user interfaces, telecommunications as well as

N. Fuhr · C.-P. Klas (✉)
University of Duisburg-Essen, Duisburg, Germany
e-mail: Klas@is.informatik.uni-duisburg.de

T. Aalberg · I. Sølvberg
Norwegian University of Science and Technology,
Trondheim, Norway

P. Hansen
Swedish Institute of Computer Science, Kista, Sweden

L. Kovács · A. Micsik
Computer and Automation Research Institute of the
Hungarian Academy of Sciences (MTA SZTAKI),
Budapest, Hungary

M. Agosti
University of Padua, Padova, Italy

G. Tsakonas · S. Kapidakis · C. Papatheodorou
Department of Archive and Library Sciences,
Ionian University, Corfu, Greece

C. Peters
CNR-ISTI, Pisa, Italy

M. Landoni
Department of Computer and Information Sciences,
University of Strathclyde, Glasgow, UK

the increasing availability of a variety of multimedia collections make it possible to offer new and better services for user groups. As DLs have been around for a few years now, an increasing number of users have some familiarity with them. The expectations and demands for better service and functionality from these users are increasing. Thus the importance of quality in DL content and services is higher than ever. However, the quality of a complex system is never better than the quality of its "weakest" component. In order to improve quality there is a need for definitions of what is intended by quality, and for the appropriate metrics and methods for evaluation.

The design and development of a DL is expensive. Evaluation results from previous systems can give guidelines as well as assist in determining methods for the construction of cost-effective and sustainable new systems. "Doing it right" in the first phases of system development is critical for the final result as the quality of an end-product depends to a great extent on the quality of early requirements and conceptual models.

The Evaluation Forum of the (2000–2003) DELOS Network of Excellence on DLs first proposed a framework for the classification of DLs, and produced the first version of a MetaLibrary of test-beds to be used in DL evaluation [1]. This was presented at the first DELOS workshop on evaluation, the Budapest workshop [2]. From the Budapest workshop two broad themes emerged: the complementary needs for metrics and test-beds, and for evaluation in the context of specific DL applications. One of the first activities of the new DELOS evaluation workpackage under the Sixth Framework Programme was to organize a workshop in Padua, Italy, October 4–5, 2004 [3]. At the Padua Workshop an extensive overview of the literature in all areas of DL evaluation was presented and different aspects were addressed: user experiences; DL models; DL usage; content evaluation; system evaluation; and experiences from the CLEF and the INEX initiatives.[1] The workshop provided a good foundation for our description of the state of the art in DL evaluation, and the identification of major issues for further research in this area.

The workshop results constitute the context, the aims and the motivation of this paper. It is recognized that DLs support specific activities in specific contexts and that they need to be evaluated to determine how useful, usable and effective they are. Therefore in Sect. 2 we introduce the principal concepts and conditions in the DL evaluation domain, and in the following section (Sect. 3) we present the state of the art in the domain. In Sect. 4, we have employed a revised version of the first DELOS evaluation framework in order to provide a model that holistically covers the current state in DL evaluation. In Sect. 5 we present a new DL framework, which operates dynamically at diverse levels. A set of recommendations is listed in Sect. 6. A summary and conclusions are given in the final section (Sect. 7).

2 Basic concepts and assumptions of DL evaluation

Although evaluation activities started soon after the first DL systems were available, the underlying assumptions and goals of these efforts were quite disparate. Thus, there is a need for agreement with respect to the central concepts, assumptions, parameters, and criteria of DL evaluations. In this section, we pinpoint some of these issues.

2.1 Evaluating the interaction between user and content

Digital libraries can be used for many reasons, but the most central set of use cases focuses around information access. Finding certain types of content, retrieving specific information, locating known items, accessing material the client does not know enough about, there are many content-based, more or less goal-directed motivations that will lead a user to the access terminal of a digital collection of information.

2.2 Current evaluation schemes

The system components that provide the user with access to the content of the library are, for the most part, based on information retrieval technology. This is a technology that is well researched and has developed its own tried and tested methodologies for evaluation. Much of what is known about the characteristics of the central components of any information system is due to a well-established and accepted formalization of the tasks they are designed to perform and a consequent evaluation of the performance. This established evaluation framework is laboratory-based and implies a fairly draconic generalization of usage, user task, and user situation. It is derived from ideas first introduced in the Cranfield experiments [4].

[1] Both the Cross-Language Evaluation Forum (CLEF) and the Initiative for the Evaluation of XML Retrieval (INEX) are sponsored by the DELOS Evaluation workpackage. See http://www.clef-campaign.org/ and http://inex.is.informatik.uni-duisburg.de/.

2.3 Current evaluation: the notion of relevance

The target notion of experimental evaluation thus far has been that of relevance. The concept of relevance is central to understanding the interactions between users, usage, information needs and informational content. Relevance – the momentary quality of a text that makes it valuable enough to read – is a function of task, text characteristics, user preferences and background, situation, tool, temporal constraints, and untold other factors. This has been understood and operationalized as a relation between documents and assessments made by a panel of human judges, in partial contradiction of the everyday use of the word.

Relevance, as measured by information retrieval evaluation cycles, does not take into account user satisfaction and pleasure, quality of information, relations between documents, or reliability of information. Most importantly, it is completely abstracted away from every conceivable context one might care to investigate. This includes the various types of contexts in which the item of information, the user, the producer, and the session may be engaged.

2.4 Current evaluation: the metrics of precision and recall

Given a target notion of relevance, system performance is evaluated using the well-accepted twin measures of recall and precision. These measures have been defined to model completeness and exactitude of systems, respectively, but not to model user satisfaction, result pertinence or system effectiveness for a given task context. Even more important, the TREC interactive track has shown that the results from batch evaluations [5] do not carry over to interactive retrieval – users are able to compensate for most of the differences observed in batch experiments.

2.5 New demands on evaluation

In DL research and deployment, new use cases and market opportunities beyond those of ad-hoc retrieval are being developed and introduced, and information access is being placed in context. When evaluating a DL as a content-bearing tool, system retrieval performance is but one component of overall system acceptability and evaluation must take more than system performance into account.

Purely laboratory-based approaches will not be able to fulfil this task. First and foremost, evaluation schemes must be designed specifically for the DL field. Secondly, the session must be taken as the unit for evaluation, not the information request. Users of information access systems interact with information, with the content of the systems they use. They retrieve material for a reason; they continuously evaluate the information they are accessing; they formulate, reformulate, and change their goals dynamically as they go; the information they access changes their perspective. They judge documents by perusing them in detail or by assessing them briefly in comparison with other available documents. An information access session is a real and perceived item in information access behavior, but systems typically are unaware of interaction on levels of abstraction higher than immediate request-retrieval turns.

2.6 Current tools and future tools – how to evaluate?

The technological support for many of these interactional aspects of information access does not yet exist other than implicitly in the system context. Most of the persistence or continuity in interaction with the system is provided by the user, rather than the system. Thus, the scope of an information retrieval component should be extended so that it is more task-effective, more personalized, or more enjoyable. In order to evaluate these systems, new metrics are required which tie together the disparate and vague notions of user satisfaction, pertinence to task, and system performance.

2.7 A new relevance?

If the concept of relevance is de-constructed and enhanced to take context into account, and information access systems are made to model information users, information producers, information access sessions, and contexts, future research will better be able to satisfy the needs of information seekers, both professional and incidental. Extending the notion of relevance so that it does not lose its attractive formal qualities is not a straightforward task, and has been attempted in various research endeavours in the past. It is completely crucial for any extension to be agreed upon by a large number of researchers. A well-crafted context sensitive relevance measure will be a durable contribution to the field, which thereafter will only be able to ignore context and usage factors at its peril.

3 State of the art

The goal of this section is to present the present the current DL foundations as well as to describe the description and evaluation trends of various research and practice sectors, which could be also adopted by the DL community.

Digital libraries represent a truly interdisciplinary research domain; existing models from a number of different communities can be applied to components of the DL paradigm. For example, the Reference Model for an Open Archival Information System (OAIS-RM) formally defines the processes for the archiving, preservation and access of digital objects [6]. The FRBR (Functional Requirements for Bibliographic Records) data schema, developed under the umbrella of IFLA, provides a framework and conceptual model for creating bibliographic records [7].

The scope of these models goes beyond DLs, and vice versa, the area of DLs is much wider than what these models cover. How broad is the definition of a DL [8,9]? So far, there is no single accepted definition. At the broadest level, the following aspects of a DL have to be considered: an infrastructure of various networked services, a structured collection of digital objects and an organization with different user roles and human interaction patterns.

Among the models for the DL community, one can find a reference model for actors and roles in DLs, proposed by a DELOS/NSF working group [10]. Within this reference model, users, professionals and agents were identified as the main classes of actors, and their roles were analyzed in characteristic DL use cases. In the context of the DELOS Network of Excellence[2] an effort is underway to complete previous efforts and to develop a comprehensive reference model for DL management systems [11].

At present a formal and general model for DLs is the 5S model [12]. According to this model a DL consists of a repository, metadata catalogues, services and a society of users. The 5S refers to *streams* and *structures* for the construction of digital objects, *spaces* for the description of digital object collections and their interrelations, *scenarios* for the definition of how services and activities change the state of the system, and finally *societies* for the interconnection of roles and activities within the user community. The 5S model is based on a mathematical formalism, and has been used in various case studies, including the generation of a taxonomy of DL terms. 5SL is a declarative language based on this model for the generation of DL applications.

Models that relate to DLs may contain or may be extended to contain a formalization of evaluation targets and aspects of quality. The 5S model has already been extended towards formalizing quality aspects of DLs. Properties such as accessibility, preservability, timeliness, and completeness have been defined mathematically using this model, so that they are usable as evaluation metrics.

Effectiveness, cost-effectiveness and the cost-benefit ratio are metrics often used for measuring performance of library services [13]. Nicholson's conceptual framework for the holistic evaluation of library services [14] proposes an evaluation matrix with dimensions of perspective (internal – library system and external – user), and topic (library system and use). This matrix helps evaluators to choose targets for measurements and methods for measuring selected targets. In Nicholson's view, evaluation implies measurement, and evaluation criteria are calculated combining some of the measurements. He introduces the evaluation viewpoints of users, library personnel and decision makers, where evaluations are refined and propagated from users through library personnel towards decision makers, and decisions and changes are propagated in the opposite direction. Optimally, he argues, changes in the library will induce measurements of their impact; the evaluation of this will suggest further changes and so on. Mapping current measurement and evaluation activities of a library onto this framework helps to relate these activities to each other and to decision making in a holistic view.

In addition to models and frameworks, there are also some practical methodologies and tools that can be used to measure the value of DL services. For example, the LibQUAL+ methodology is a complete market survey of user perceptions for identifying gaps in service delivery [15]. It is used extensively in the U.S. within the Association of Research Libraries (ARL) and it is under continuous development in pursuing of applicability in the DL sector [16]. The goal of eVALUEd[3] is to produce a transferable model for e-library evaluation and to provide training and dissemination in e-library evaluation.

These models differ from DL evaluation models, due to the relation of the DL operation outcome with certain served communities, for example impact or satisfaction of geographically fixed communities. The outcome of this section is the inherent difficulty in evaluation activities. The complexity of DL constituting elements, structure and provision determines to a great extend the variety of evaluation activities in terms of formalism, dimensions, depth and practicality. This was reinforced by the outcomes of the first survey carried by the DELOS WG on evaluation, mainly in the area of semantics and constructs granularity [1]. It is a notable observation that under current conditions evaluation in whole is a

[2] See http://www.delos.info/.

[3] http://www.evalued.uce.ac.uk/.

multi-faceted process that demands contributions from varied and distant in between communities.

4 Defining a conceptual model for DL evaluation

The DELOS Working Group on Evaluation has worked on the definition of a DL conceptual model. The initial version of the model [1] was discussed at the first DELOS evaluation workshop held in Budapest, 2002.

This approach (see Fig. 1) uses a generic definition of a DL 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. The definition of the set of users predetermines the appropriate range and content of the collection (thick arrow connecting "users" to "collection"). The nature of the collection predetermines the range of appropriate technologies that can be used (thick arrow from "collection" to "technology"). The attractiveness of the collection to the users and the ease with which they use the technologies will determine the extent of usage of the DL (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).

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 principal 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. Evaluation criteria are assigned to each of the components, and metrics are suggested for these criteria.

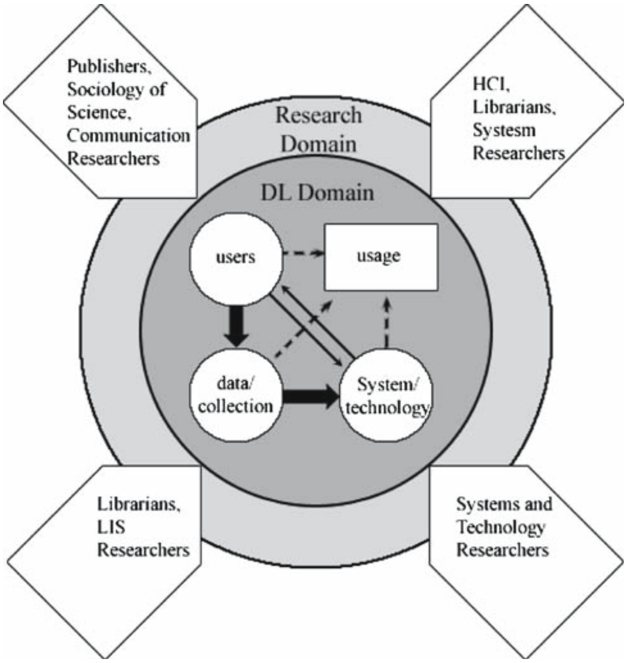
Publishers, Sociology of Science, Communication Researchers
HCI, Librarians, Systesm Researchers
Research Domain
DL Domain
users
usage
data/ collection
System/ technology
Librarians, LIS Researchers
Systems and Technology Researchers

Fig. 1 A DL classification and evaluation scheme

4.1 The interaction triptych model

At the second DL evaluation workshop in Padua [3], special attention was given to the relations between the components of a DL, i.e. the relations User–Content, Content–System and User–System [17], where user here represents both the user and the usage nodes of Fig. 1. This is shown in Fig. 2. The Content–System pair is related to the performance attributes (precision, recall, response time, etc.), the User–System pair is related to usability aspects (effectiveness, satisfaction, etc.), and the User–Content pair is related to usefulness aspects. These pairs represent the relations – interactions among the DL components and define a three-axes framework for the DL evaluation which we call Interaction Triptych [18]. The following subsections are structured according to this logical model when discussing various evaluation aspects.

4.1.1 User

Users is the first component of any interaction process and his characteristics are complex and constantly evolving [1]. Even in specific evaluation projects, where research parameters are constrained, the number and the attributes of the users are many and complicated.

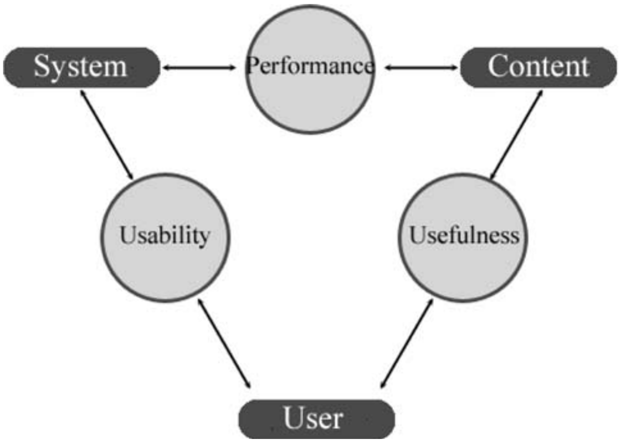

Fig. 2 The interaction triptych model

Whatever the particular methodology chosen, when performing a user-oriented evaluation, the objectives are to acquire a deep understanding of the user requirements for technology design and planning, and to receive systematic user feedback throughout the design process. Marchionini [19] evidences two important issues that must be taken into account: evaluation research should be longitudinal in order to capture a rich and reliable data set for analysis, and it should be multifaceted using a combination of methods. Each evaluation may also have its own set of measures and data collection methods.

4.1.2 Content

Content is the prime reason for interacting with a DL. This component addresses the user's information needs. The relation between the user and the content strongly depends on the informational need of the user. The perceived usefulness of content is the first selection criterion for the user. During a search session, the users' perceptions of the usefulness of the content and their information needs can change and they may be forced to reformulate them and re-direct the information strategy (if any).

Content characterizes the DL and affects the consequent processes dramatically. It is indispensably bound with the purposes of the DL and the desired outcomes of its operation and serves the information needs of a community. The evaluation of content must take place under strict rules and standards that guarantee user access to high quality, appropriate-for-their-needs information. However, underneath this user-centered layer of evaluation lie numerous other issues; issues related to the nature, the structure and the administration of the content.

4.1.3 System

The *system* is the best known component of the interaction process, as it is governed by the rationale of the developer. It consists of various subsystems that perform different basic and auxiliary operations. DL systems generally include collections of multimedia digitized data and services that help storage, access, retrieval and analysis of the data collections. The final aim of a DL system should be that of enabling people to access human knowledge any time and anywhere, in a friendly multi-modal way, by overcoming barriers of distance, language and culture, and by using multiple network-connected devices [20].

In order to fulfill all the requirements on it, a DL system must be the result of the integration of a number of different components. Discussions as how to implement each single component depends on the architectural choices, on the type of data access and retrieval, on the visualization and personalization of information. The individual evaluation of such components can be carried out following evaluation standards already available in the literature. However, an important issue is the evaluation of how the individual components interoperate inside a DL. The evaluation of a DL system makes sense mainly if it is performed in relation with users and contents.

4.2 Axes of evaluation

As stated, the evaluation process should be based on the relations of the DL components. These interactions define our Interaction Triptych which consists of the following axes:

- *Usability*, which defines the quality of interaction between the "User" and the "System". This helps the user to manipulate a system effectively, in an efficient and enjoyable way and to exploit all the available functionalities. A usable system is easy to learn, flexible and adapts to user preferences and skills.
- *Usefulness*, which concerns the "User" and "Content" components. The usefulness of the content and its relevance to the user tasks and needs are the reasons behind the selection and usage of a DL. This relevance is translated into actual relevance, type and level of resource relevance and task relevance.
- *Performance*, which is placed between the "Content" and the "System". Although it is of crucial importance, this is an aspect of the system that the user cannot see or evaluate directly. The performance of the system depends strongly on the formats, structures and representations of the content.

4.2.1 Usability and related studies (User–System)

In our conceptual model, usability constitutes a major research field that derives from the interaction between the user and the system. It represents a set of principles that should ensure that the system helps the users to conduct their tasks in an efficient, effective and satisfactory way. According to the most commonly quoted definition "usability is the extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use" [21]. Usability studies not only the surface representation of a system, but also its underlying structures. A usable system needs to be attractive to the user, but above all, must be learnable, safe and reliable. Usable systems need to be error tolerant, consistent, adaptive to the terminological and operational preferences of the users, supportive, easy to be remembered and visually clear. Usability is not only an issue for the DL research area. A major part of the market sections of DLs and electronic publishing is investing a considerable amount of money in the application of usability principles and methods in order to produce usable large-scale systems [22].

Blandford, et al. [23], in the editorial of a special issue of the International Journal of DLs on the usability of DLs, underline the danger of producing DLs with novel and advanced features, which will be neglected by the users due to the difficulty of using, learning and understanding their operation. The wealth of research into the usability of library websites has uncovered the syndrome of "librarians know better" [24]. This syndrome can be found everywhere and it means that the developers are ignoring the users' primary needs, behavioral models, operational skills and context infrastructure.

Usability evaluation is conceived as "the act of measuring (or identifying potential issues affecting) usability attributes of a system or device with respect to particular users, performing particular tasks in a particular context" [25]. In this wide sense evaluation is conceived as all those processes that (at least) describe and evaluate the impact of specific interface features on the creation of a normal and effective interaction between the user and the system. In any case, usability is the main key for system developers to address user centered design issues.

Usability issues are located at different stages of the development cycle. Formative and summative evaluation share common techniques but have different perspectives and the results vary with respect to the number and weight of recommendations. System developers apply a wealth of techniques that range from fully automatic ones to techniques that necessitate the involvement of real users in controlled environments or operational contexts. Automated evaluation methods have been criticized for their lack of ability to produce interpretable and qualitative data that will help the evaluators understand the nature of the usability problem, its severity and the impact it has on the user interaction [26]. Techniques based on user participation are described as time and resource consuming. Nevertheless, the wealth of usability evaluation techniques can be grouped into five main categories:

- *Automated techniques* that use specific usability evaluation software to produce mostly quantitative data. In this category we can include transaction log analysis, a widely used technique that examines the activity of users in a given time session [27–29].
- *Empirical or research based techniques* that require users' participation, observation and recording of their actions [30–32].
- *Comparative techniques* that collate the system with given standards and guidelines or other operating systems. This may be performed automatically.
- *Text interviews, focus groups, surveys and questionnaires* are the prime methods for collecting qualitative data, like user preferences, explanatory opinions and satisfaction rates [33,34]. Multiple variations exist for each of these techniques, and also for the way in which they are conducted (e.g.: online, in person) and the level (e.g.: structured, semi-structured interviews).
- *Analysis techniques* that require the participation of usability experts. They can examine the system, either alone, or with the cooperation of domain experts for better results [35].

CALIMERA, a project funded by European Commission's IST Programme, has done a significant amount of work on usability [36]. Dicks, quoting Rubin, notes that usability evaluation has inherent limitations [37]. One of these is the fact that the testing procedure is an artificial situation. Thus the results of a usability testing will not be fully representative. The HCI community has struggled to lift these inherent limitations by recommending the combined application of several methods in a sequential and repetitive manner [38]. In the DL research field this recommendation has been widely used, see for example the ADL [39], DeLIver [40] and Envision [41] projects. The combination of different methods has taken the form of the "*convergent paradigm*", which, apart from improving system usability, can lead to comparative evaluation of the methods themselves [42].

Other ways of minimizing the effects of usability testing exist, such as letting the users perform their own tasks [43] or developing and practicing techniques that utilize users' spontaneous feedback during real time interaction [44].

User interface related studies Generally, the user interface acts as a bridge between the user and the system-environment so that the user can interact with the DL system. It is also the means by which the system reveals itself to the user. A DL must provide support for interaction. In fact, it must support different types of interactions and therefore appropriate evaluation methodologies must be developed to measure the effectiveness of these interactions.

In a DL, the complexity of the interactions is determined by the dimensions of the users, the information sources and the objects as well as the complexity of tasks to be performed. Some time ago, Bates [45] suggested a four-level approach to search activities (move, tactics, stratagems, and strategies) that needed to be supported by an information access system (such as a DL) in order for a user to retrieve the information needed. Her proposal clearly refers to interface issues as well as to task-related issues. Cousins [46] suggested an early solution to the problem by proposing a task-oriented interface for a DL. This interface would need to take into account a range of activities and processes, and in particular the variety of:

- information sources,
- information objects,
- users and groups of users.

Any evaluation of the user interface also needs to take into account how well the different components and functionalities are integrated. This will affect the results of the information handling task.

Task and context-oriented studies Each domain and context contains a specific set of components that form that particular DL environment. The contexts in which specific information access situations occur are therefore important. One of the goals of a DL will be to support users and communities in their task performance processes. Criteria that can be used for evaluation need to be indicated. An example of the group of studies that is concerned with work-task and domain related evaluation studies is the Cognitive Framework for Analysis and Evaluation [47]. Evaluation is a guided bottom-up approach that focuses on individual as well as collaborating users and the context of their activities within the means and ends of their work domains [48].

Within a task-oriented evaluation perspective, aspects such as interaction and interactivity should be considered. Aspects that can be evaluated include:

- Types of tasks and stages of a single task.
- Task performance and task procedures.
- Task complexity and task variations.
- Differences between domains.
- Impact on organizational levels.
- Types of information access systems and processing activities.

Usage studies Aspects of the evaluation continuum that have not been sufficiently evidenced regard use and usage. When dealing with different types of information systems, the system design usually ends when the user has retrieved the required information. But the information handling process does not end there. There are other issues to be examined like: What part of the retrieved information is actually used and in what way? When and how is the acquired/accessed/retrieved information used? What types of information components are used and what type of outcome is produced.

As Borgman [49] points out, creating, using and seeking information involves a continuing and iterative evaluation process, which may involve a variation of evaluation methods and measures. The choice between the various methods is in itself a research question: the last word on evaluation methodologies has not yet been written, neither with regard to the art of methodology development nor to the craft of methodology choice for practical evaluation purposes.

4.2.2 Usefulness (User–Content)

Usefulness is the abstraction of every information need that stimulates user interaction with DLs. It lies between the content and the user needs and it reflects how users perceive the relevance of a DL with their needs, the width, the breadth, the quality, as well as the validity of its collection, and the ability to serve their goals. The assessment of usefulness depends on the features of both the user and the content components. User monitoring provides us with an insight into the way the user seeks information and uses it in the DL environment. Two major approaches are followed when addressing these issues, namely user studies and information behavior. These are two multidisciplinary areas that can address different variables of the evaluation process. Either as main evaluation processes, or as supplementary components (on a priori and/or posteriori stage), they provide valuable tools for assessing the usefulness of the information content.

User studies An evaluation of DL collections and services must take into account the characteristics of the user community. User-oriented evaluations are well known and have been applied in many studies. The evaluation of a DL can serve many purposes including understanding phenomena such as human information seeking behavior and information handling and refinement. User-oriented evaluation involves many stakeholders, both individual end-users and librarians, as well as various groups from the community or society in general. Criteria used in the evaluation of user–system interaction include:

- Types of users and their characteristics, such as different levels of knowledge and experience.
- The information needs of the users.
- Different satisfaction factors (e.g.: functionalities and task accomplishment).
- Types of information handling strategies.
- Tasks and task procedures.
- Representation of work domains and environment.
- Collaboration between users and groups of users.

User studies are the most common means to collect and assess information about the preferences of a user community. Surveying techniques are used for the collection of data and to produce statistical representations of the preferred properties of the DL content, the way and the modes of access.

Moreover differences between classes of users are revealed. Individual interests as well as dissimilarities between user classes constitute informative estimators on the generalized value of the DL. Some of the user studies mentioned in the 2003 CLIR report, by Tenopir [50], focus on the profile generation of users who use electronic resources in a physical environment. We still need to form a clear image of those users who do not enjoy the advantages of a physical organization that could help them solve problems encountered when accessing the DL.

As mentioned earlier, user studies exploit surveying techniques, like questionnaires, surveys and online forms [51]. In the case of small samples more direct methods are selected, such as interviews and focus groups. Direct contacts between the evaluator and the participant help the former to interpret previously collected statistical data, to explain behaviors and to check user concepts [52]. As already mentioned, another practical method of data collection is the employment of transaction logs. This makes it possible to collect and analyze all traces of a user's choices during a work session and thus create meaningful models of usage. Despite criticisms of the cohesion of transaction log analysis on the web with cached files, this method is widely used in many evaluation projects around the digital information globe. The Super Journal project investigated the employment of transaction logs for the extraction of useful behavioral models and then developed an evaluation framework for the selection of the appropriate data for collection [53]. A coordinated effort among some DELOS partners and TEL has been launched in 2006 to conduct search engine and Web server log analysis to gain insights for improving user–system usability in the context of a large DL such as TEL is.[4]

Information behavior This research field focuses on the patterns of behavior that are generated between different classes of users and different information systems. Information behavior is an interdisciplinary field that involves personnel, methods and applications from psychology and cognitive science, computer science, and information science [54]. Information behavior investigates the information seeking and retrieval process in detail. It splits this process into separate stages and analyzes them in correlation with actions, cognitive states and emotions. Interesting research issues are the motivations of usage, the affective states of the process [55], differences between certain classes of users [56], the structured or unstructured nature of accessing information, the serendipitous selection of information seeking directions [57], and the problematic states [58]. Methods for collecting data in information behavior studies are once again surveying techniques and questionnaires, as well as interviews [59] and focus groups. Finally observational studies, including recording of user interaction and think-aloud protocols, are also used.

Content-related studies The following issues may affect DL policy and decisions, as well as the technology infrastructure and processes (e.g. server technology, maintenance). In an interdependent environment, such as a DL, evaluation cannot study these factors in isolation; they must be measured in the context of the entire system. A clear picture of the properties of the content is an essential part for the DL evaluation on the axis of usefulness. Apart from acquiring a user population profile, we also need to know what kind of content is available and how access and adoption is affected by it. The following issues represent the major research areas regarding the properties of content.

- *Content nature*
  - Form (text, video, images, sounds)
  - Language (language(s) used and whether the content is monolingual, bilingual, multilingual)
  - Method of creation (digitized, born digital)
  - Type (white, grey literature)
- *Content structure*
  - Level (primary object, metadata)
  - Size
  - Cohesion (homogeneous or heterogeneous attributes usually apply at collection level and are

[4] See http://www.TheEuropeanLibrary.org/.

connected with interoperability standards and technologies)
  - Delivery mode (hypertext protocols, streaming technologies)
- *Administration*
  - Collection growth (acquisitions, drop-outs)
  - Rights (related to the license mode and content type, e.g. evaluating policies on thesis acquisition and provision)
  - License mode (open access, one-off purchasing, subscription-based with multiple variations such as moving walls).

It is obvious that some of these issues are easily resolved (for example size), whereas others need specific tools and applications, and may be related to legislative parameters. Transaction log analysis may reveal the rate of preference on each of the content attributes, search engine and Web server log analysis can serve to inform on preference of the users and to personalize services for the users themselves, while other methods, like Conspectus [60] can assess associations between collections and inform about the collection growth.

4.2.3 Performance (System–Content)

Given a general framework of evaluation (or context as in [61]), in this section we mainly concentrate on a system-centered approach. This approach is one of the most prevalent, and involves a study of performance aspects. It also includes the assessment of the efficiency of a feature, a specific design, or a technological component. With respect to the classification scheme of a DL given in [1], we focus on the System/Technology view of a DL. In particular, we concentrate on the evaluation aspects of information access and extraction.

In general, the evaluation of information systems has been carried out in controlled settings, using different criteria and measures, with the aim of providing benchmarks and improving performance. The performance evaluation of a DL is a non-trivial issue that should cover various aspects, such as: architecture, storage and access capability, and management of multimedia content. Although DLs by their nature are distributed modular systems in which various components cooperate together using some network capabilities, evaluation campaigns organized to evaluate DL components rarely exploit this distributed nature. In this section, we discuss important issues regarding both classic "off-line" evaluation and also some of the questions raised by the emerging areas of distributed access/computing ("on-line" evaluation).

"*Off-line*" *evaluation* Current evaluation campaigns present the participants with one or more test-bed collections, a set of tasks to be performed, and a method by which the performance of their systems can be evaluated with respect to the collections and the tasks. This kind of approach is found in the most important evaluation campaigns such as: TREC,[5] CLEF,[6] NTCIR[7] and INEX.[8] However, a more mature way of evaluating performance within these campaigns would also take into consideration the following aspects:

- Every year collections change constantly. The number of different collections usually increases as well as the total amount of data to be processed. This constant growth can be seen in relationship with the capacity of computers to store even more bytes. However, comparing system performance in different years is almost impossible. Therefore we need methods to make performance analysis comparable over time.
- Together with the question of how to make collections stable, there is also a problem of a lack of a historical/temporal vision of system evaluation. Deciding whether a maximum level of performance has been achieved for some specific feature/design/component or not is a non-trivial question. Even if the focus is on one single aspect of a DL, it is not easy to claim that an asymptotic limit of best performance has been reached or improvements can still be made.

"*On-line*" *evaluation* In the campaigns mentioned in the previous paragraph the search mechanisms of DLs are evaluated independently of other architectural features. In [62] aspects of how to manage the distributed nature of DLs were explored in depth. It is clear that new areas like peer-to-peer (P2P), grid computing, and service-oriented computing provide new opportunities for DLs. For example, data management by means of P2P architectures allows for loosely coupled integration of information services and sharing of information such as recommendations and annotations [63]. DLs where multimedia content is a key-point (such as protein structure display in biomedical DLs), and an intensive computational load is required, can use a grid computing middleware. Other issues such as describing the semantics and the usage of information services and combining services

[5] http://trec.nist.gov/.

[6] http://www.clef-campaign.org/.

[7] http://research.nii.ac.jp/ntcir/.

[8] http://inex.is.informatik.uni-duisburg.de/.

into work-flow processes for sophisticated search and maintenance of dependencies can be provided by service-oriented computing procedures. This intrinsic feature of a DL suggests that the evaluation process should also be in a sense on-line, meaning that the evaluation process should exploit the network and the modularity of a DL.

4.3 Practical issues

In the previous subsections we outlined the components of a DL and the interactions between them, as well as the principal directions, methods and techniques for their evaluation. The following questions refer to the practical implementation of an evaluation procedure. As in every evaluation project, the selection of the appropriate method or technique is determined by some, often inter-dependent, factors. Some of these concern:

- Cost (how much will it cost in terms of money, resources and time?)
- Effectiveness (how effective will it be/how many problems will be revealed?)
- Time (how much time is required?)
- Personnel (are skilled staff or usability experts needed?)
- Infrastructure (what kind of storage and analysis software is needed?)
- Place (can the evaluation be done remotely/can it be done in controlled environments or in contextual conditions?)
- Evaluation timing (at which stage of the development cycle will it occur?)
- Pace of evaluation (how frequently will it take place, with what aims and under which circumstances?)
- Collaborative evaluation (are test-bed systems needed for the evaluation?)

One major aspect that will affect evaluation deals with the type of experimental framework in which the evaluation will take place [64]. We analyze the factor “Place” since DLs have to serve physically remote users in diverse time periods. This requirement will impact on the criteria, measures, data collection methods, procedure and outcome. Generally, we apply either a laboratory-based evaluation or a real-life empirical evaluation setting. Usually, there is a dichotomous relationship between these approaches but, in fact, they could complement and support each other at a higher level of evaluation. For this reason, it would be interesting to be able to integrate statistical data and narratives in order to assess impact as well as performance and usage. One of the differences between these two approaches is the extent to which the evaluation process and outcome are controlled. In an experimental, laboratory-based setting we can control what and how we evaluate, whereas in a real-life situation there can be components of the evaluation setting that we cannot really control and therefore different sets of data collection methods may have to be applied.

The process of collecting qualitative data through remote methods is very difficult; so far we have insufficient evidence with respect to several issues, e.g. how to recruit and stimulate user participation in remote evaluation settings. The DLF/CLIR 2002 report on usage and usability assessment indicates a number of concerns about assessment methods, for example the need to collect meaningful and purposeful data, the development of skills to gather, analyze and present user data [65]. DLs are systems that, even when they have been developed to serve physically located user communities, still need to provide a remote and asynchronous mode of access to information. Thus we need methods that effectively evaluate them for cases of full remote access, where the user does not have any physical connection with the DL administration.

The selection of the right evaluation method is dependent on the above mentioned factors and their combination with the type of data that will be collected. Methods that involve qualitative data collection include think-aloud protocols, observation, interviews, questionnaires and simulations, while, for quantitative data collection, methods such as transaction log analysis, and error and time analysis can be used.

Digital libraries are destined to serve user communities [66]. If unused these systems fall into oblivion and terminate their operation. This fundamental commitment of DLs determines an evaluation of the role that the DL plays in serving a community and its members. The difficult task of this evaluation is to assess in an effective way the qualitative serving of this community and its members. There are diverse levels of assessment. In addition to the level of evaluation (individual/community), other factors like the mode of use (e.g. collaborative) and the context of use (information access, education, research) make the whole process even more complex. System developers and evaluators must implement the right evaluation process, so that it is more efficient and can generate useful system design recommendations. Therefore a prerequisite of this type of evaluation campaign should be the careful and analytical mapping of the evaluation research onto the real needs and aims. This process must be done in a systematic way in order to cover all facets, to address the main questions and should be finalized in a set of methods, criteria and metrics that will be used in the campaign.

4.4 A classification model of evaluation

Evaluation Computer model [67], given in Fig. 3 is an analytic mechanism of description and classification of evaluation procedures. Based on a faceted analysis, the Evaluation Computer is a framework that can be used to create comprehensive profiles that could be comparatively listed in order to assess the depth and width of coverage of the current evaluation foci. According to this model the evaluation process is defined as a selection of a set of points in a multidimensional space. The space consists of a set of aspects, namely evaluation, user, organizational, content and system. Each aspect is divided in a set of facets. Thus the Evaluation Computer is able to systematically record the main items regarding the evaluation process (evaluation aspects), the DL components (user, content, system aspects) and the contextual conditions (organizational aspects).

The central section in Fig. 3 provides an example. The first part of the evaluation profile building process (the inner ring) concentrates on the evaluation aspect. In the case of this example, it indicates that the nature of the evaluation activity is summative and that the method selected is transaction log analysis. The second part of the profile (the next ring) is determined by user aspects. The evaluation activity here uses logs resulting from usage by the general public, and there are no geographical constraints. This worldwide user community uses the DL for educational purposes (organizational aspects). What is evaluated is shown in the fourth ring: the interaction of users with the content and more specifically with textual objects in their full format (content aspects). As shown in the outermost ring, the users interact with the DL through a web interface, while the distribution parameter is based on the dominant "client-server" architecture (system aspects).

This example illustrates the ability of the Evaluation Computer to classify an evaluation activity, both at the process and at the DL component level. The resulting classification can be compared against the existing corpus of evaluation studies. This kind of analysis could help designers to better understand the reasons underlying particular decisions in previous studies, the final results and implications, and the relation between the results and the undergoing activity.

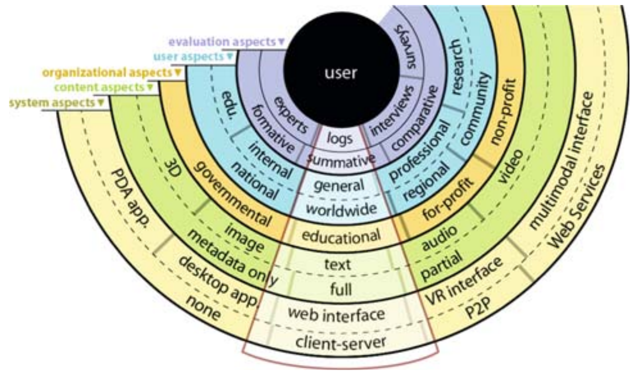

Fig. 3 A possible layout of the evaluation computer

5 DL evaluation framework

It is clear at this point that setting up a framework for the evaluation of DLs is not only a very difficult task but also quite controversial with so many models and formalisms already available and so many contributions from related research areas.

We base our work on the framework for DL evaluation designed by Saracevic [61]; he introduced four dimensions or components (construct, context, criteria and methodology) for describing evaluation activities, which we describe in more detail in the following. The framework per se acts both as a classification scheme for existing evaluation studies and a model for new ones. In particular, a set of guidelines can be extracted from it in order to make the overall evaluation experience not only less controversial and complex but, more importantly, easier to replicate and compare to similar ones so that, in the future, evaluation experiments and their findings would not be obsolete individualistic exercises but would serve the entire community and allow for progress and improvements where needed. Saracevic's four dimensions have also proved to be naturally descriptive and sufficiently flexible to describe a variety of studies. This is an essential feature if the framework is to be widely adopted by the DL community. As a whole, they respond to the most crucial questions raised within the DELOS Evaluation forum:

- Why evaluate?
- What to evaluate?
- How to evaluate?

The *why*, *what* and *how* questions are immediate and intuitive to use and understand, but clearly have some overlaps in terms of their coverage. Furthermore, taken in isolation, they are open to individual interpretations and their complexity can grow out of control. This could make their application to model existing evaluation initiatives or design new ones cumbersome, complex and inconsistent.

Saracevic's model has instead proved its usefulness as it provides four clean cut categories while still maintaining the rich expressivity necessary for this type of framework [68–70]. By using a more concrete terminology, specific to evaluation, it is easier to classify existing studies keeping in mind the important questions.

5.1 The four dimensions

Construct represents what is being evaluated and how it can be used both at a high and a low level. For instance, it could be applied in cases where the evaluation team defines its own interpretation of the DL. Similarly it can define those aspects and/or components of the DL that are objects of the evaluation procedure. Construct answers the WHAT question at as many levels as needed.

Context is possibly the richest of all dimensions as it accounts for everything that qualifies as motivation and framework for each evaluation study and as such covers scenarios, actors, objectives and goals, approaches and perspective in the study. It could be further developed into additional dimensions in order to provide all the necessary elements to describe the framework for each individual evaluation study. This is where the WHY question is answered.

Criteria is actually the core of the evaluation study and covers parameters, factors and measures used to assess the quality of what is evaluated and every aspect of a DL being evaluated. It responds partially to the HOW question.

Methodology together with criteria also answers the HOW question and provides a means to describe the procedures and protocols followed in both the gathering and the analysis of data from the evaluation experiment.

None of the four dimensions is independent of the others but, taken as a whole, they provide a self-contained environment where the various issues related to DL evaluation can be explored and described without the risk of unnecessary overlap.

5.2 Evaluation models – Contributions to the framework

Although the three models possess a significant descriptive, planning and developmental strength, they are placed under the umbrella of the Saracevic model. The Saracevic model defines the wide and narrow environmental conditions of the evaluation initiative and, as it is based on a comprehensive review of the evaluation literature, encapsulates the knowledge and experience of varied disciplines and applications in an abstracted fashion. The models are placed under its umbrella, as they share similarities with its taxonomic properties, as well as they have a common mentality in their construction. Table 1 demonstrates the contribution of the three models used for the development of a conceptual DL evaluation model in the formation of a new frame-

Table 1 Contributions of models

	DELOS WG model	Interaction Triptych Framework	Evaluation Computer
General Aim	To classify and present an evaluation schema. It identifies main entities in the DL lifecycle on a deterministic route.	To identify DL main entities, trace their relations and to define evaluation axes and criteria under the user interaction point of view.	To formulate evaluation profiles and encourage comparisons between them. It serves as a comparison benchmark.
Why *Context*	It defines four components describing DL operation (users, usage, data, system) and addresses system-centered evaluation aims.	It considers DL as an interaction of three entities (users, content, system). It considers that these entities define three evaluation axes (usefulness, usability, performance). The evaluation focuses on the user interaction improvement.	An evaluation profile is considered as the application of an evaluation method to a DL contextual instance. A DL contextual instance is a tetrad of values from the domains: user, content, organization, system. Each evaluation profile constitutes a complete evaluation process and its results could be compared with the results of other profiles.
What *Construct*	It suggests the development of a testbed that aggregates the requirements for DL operation and evaluation.	It considers evaluation as a composite process analyzed to three axes (usefulness, usability, performance).	It analyzes DL operation to four main components (user, content, organization, system). It combines different evaluation methods with the four components defining particular evaluation profiles.
How *Criteria*	As a generic model it encourages the researchers to define particular criteria for each evaluation experiment.	For each evaluation axis it provides a particular and rich set of criteria and metrics.	It provides a mechanism (evaluation computer) for the generation of evaluation profiles.

work. Two issues constitute the difference between these models:

- Orientation: DELOS WG is system-centered, ITF is user-centered, while EC is neutral;
- Depth of provided tools: vertical-horizontal. Furthermore, while the first two models propose evaluation entities in the broad and narrow view of DL operation, EC provides an overview of the DL evaluation set-up.

5.3 Proposed guidelines – how to adopt the framework in the most productive way

The idea behind adopting this high level framework is to provide designers and evaluators with a better understanding of the role and the benefits of a well designed and structured evaluation experiment. It is also a way to support evaluators in setting up effective experiments, so that their findings can be compared with similar or related studies and thus maximize their utility. The four dimensions can then be used in a variety of ways, according to the level of development of the evaluation and the needs of its designers. In a very initial stage, in preparation for the design of the evaluation study, these four aspects can be used to group similar case studies into clusters, so that the evaluation designers can obtain a better understanding of what the rest of the community is focusing on and what has been neglected.

The next stage for the evaluators is to decide on which aspects to focus according to their needs and the project requirements. Here the four dimensions are used as the basis for a brainstorming process in which designers elaborate and adapt them to their evaluation needs. They help to structure the answers to the crucial questions: why, what and how.

- Why evaluate; i.e. determine the aims of evaluation. In this stage strategic decisions are taken regarding the constructs, the relationships and the evaluation itself. For example, how satisfied are the users of a DL (construct: user) with the effectiveness of an existing IR system (association: user - system) and what results do the evaluators expect to be produced (e.g. a summary of system deficiencies in comparison with design aims). The Evaluation Computer could be used to formulate a profile of the evaluation process itself through the "Evaluation Aspects" ring.
- What to evaluate; this involves:
  - Determining the constructs (components, relationships)
  - Determining the type (comparative, formative, summative)
  - Determining the target DL service.

  The framework presented at the DELOS Padua Workshop [17] and shown in Fig. 2 indicated the main constructs of an evaluation procedure, as well as their associations. The interaction axes represent three main evaluation foci, which are discussed extensively in the literature. In conjunction with the Evaluation Computer, this framework can be used to define the main constructs and sort their attributes (e.g. dividing content attributes by semantic and morphologic properties).
- How to evaluate; i.e. decide on the way to perform the evaluation:
  - Planning the evaluation by selecting methods, criteria, metrics, samples (e.g. humans, collections)
  - Executing the evaluation by selecting and analyzing data (main methods and alternatives)
  - Presenting the results.

At this point, the process materializes all the decisions taken in the previous stages. The Interaction Triptych model described in Sects. 4.1–4.2 presents the main evaluation criteria and metrics mapped onto the three interaction axes. This could serve as a starting point for the evaluators to find the appropriate measurements and to adapt them to the needs and requirements of their research.

Once the evaluation experiment has been designed and the operational procedure is defined, the four dimensions can be used for the analysis of the data collected. This has the twofold advantage of providing direction to the evaluation and providing results comparable to those from other studies inside the same framework.

It is worth noting that any tool of this level of expressiveness and effectiveness is valuable to a community as long as it is easy to use and flexible to adapt. The proposed framework has both qualities. The verticality of the four dimensions permits the incorporation of various evaluation frameworks, the analysis of the current conditions and the proposal of solutions, as has been illustrated in the case of the Evaluation Computer and Interaction Triptych frameworks.

Overall the four dimensions can be used in a number of practical and effective ways during the design, implementation and final analysis of any evaluation study. DLs can be complex systems and their evaluation is a challenging exercise. Any tool that could help designers in this task should be carefully considered and adopted by the DL community if it fulfils their needs. In our opinion, the simple framework proposed above could have a positive impact in a variety of ways by assisting and supporting designers of DL evaluation studies at different stages.

6 Recommendations

A number of recommendations emerge from the multi-faceted analysis presented in this paper and should be taken into consideration when setting up a DL evaluation activity:

Flexible evaluation frameworks For complex entities such as DLs, the evaluation framework should be flexible, allowing for multi-level evaluations (e.g. by following the six levels proposed by Saracevic [61], including user and social). Furthermore any evaluation framework should undergo a period of discussion, revision and validation by the DL community before being widely adopted. Flexibility would help researchers to evade obsolete studies, that are based on rigid frameworks, but to use models that can "expand" or "collapse" at their project's requirements and conditions.

Involvement of practitioners and real users Practitioners have a wealth of experience and domain-related knowledge that is often neglected. Better communication and definition of common terminology, aims and objectives could establish a framework of cooperation and boost this research area.

Build on past experiences of large evaluation initiatives Evaluation initiatives, such as TREC, CLEF, INEX, and NTCIR, have collected a wealth of knowledge about evaluation methodology.

In order to foster evaluation research in general, the following issues should be addressed:

Community building in evaluation research The lack of globally accepted abstract evaluation models and methodologies can be counterbalanced by collecting, publishing and analyzing current research activities. Maintaining an updated inventory of evaluation activities and their interrelations can help to define good practice in the field and to help the research community to reach to a consensus.

Establishment of primary data repositories The provision of open access to primary evaluation data (e.g. transaction logs, surveys, monitored events) as is common in other research fields, should be a goal. In this respect, methods to render anonymous the primary data must be adopted, as privacy is of strong concern. Common repositories and infrastructures for storing primary and pre-processed data are proposed along with the collaborative formation of evaluation best practices, and modular building blocks to be used in evaluation activities.

Standardized logging format Further use and dissemination of common logging standards is also considered useful [71]. Logging could be extended to include user behavior and system internal activity as well in order to support the personalization and intelligent user interface design processes.

Evaluation of user behavior in-the-large Currently, evaluation is focused too much on user interface and system issues. The user satisfaction with respect to how far his/her information needs have been satisfied (i.e. information access) must be investigated, independently of the methods used to fulfil these needs. The determination of user strategies and tactics is also recommended (such as search strategies, browsing behaviors). This relates to evaluation in context, and to the question of identifying dependencies in various contexts (e.g. sociological, business, institutional). Collecting user behavior as implicit rating information can also be used to establish collaborative filtering services in DL environments.

Differentia specifica of the domain of evaluation An important problem is how to relate a possible model of DLs to other overlapping models in other areas. How does a DL relate to other complex networked information systems (e.g. archives, portals, knowledge bases, etc.) and their models? Is it possible to connect or integrate DL models to the multitude of related existing models? The answer to this question should also help to define the independent research area of DL evaluation.

7 Conclusions

In this paper we described the state of art in the DL evaluation research area. In parallel we have attempted to establish a framework for evaluating DL systems in a holistic way. Our objective is to formulate a framework that could accommodate as many aspects as can be found in the various levels of the evaluation procedure. The multitude of evaluation foci and the variety of perspectives have been a concern during the formation of the framework. Finally, we have provided recommendations concerning the actions needed and the strategies to be followed in DL evaluation. It is expected that the unification of the multi-form and varied DL evaluation activities within a common framework will need considerable time as DL research and development is still at an early stage and wide-scale awareness, acceptance and employment of DL systems is only just beginning. However, evaluation is and will remain a crucial aspect in the evolution and acceptance of these systems.

Acknowledgements The work has been supported by the DELOS Network of Excellence on DLs, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

Evaluation of Digital Libraries: Used Terminology

Assessment: The process of measuring, quantifying, and/or describing the attributes of a system covered by the evaluation

Evaluation: The systematic process of determining the merit, value, and worth of something. Evaluation is broader than assessment and involves making a judgement as to the effectiveness of an assessment. Evaluation has a certain goal, methodology and devices or techniques. The goal suggests some evaluation criteria (e.g. performance), which sometimes break down to evaluation parameters (e.g. response time, availability). At times these parameters are qualitative, at times quantitative. If quantitative, a metric can be defined. In the case of metrics the values are of course comparable over different evaluations. A successful evaluation should be transferable (apply to other contexts) and confirmable (verifiable).

Evaluation, comparative: An evaluation activity that denotes the degree or grade by which a system or component has a property or quality greater or less than that of another.

Evaluation, formative: An evaluation activity carried out in parallel with the development phases of a DL system. As a part of development, a formative evaluation aims to minimize system imperfections before release.

Evaluation, summative: An evaluation activity carried out at the final stages of the development cycle or after an initial release in order to measure the performance in real operating conditions.

Evaluation device: Instrument or application that is used as part of an evaluation process.

Evaluation framework: a defined structure of methods and supporting devices designed to support the evaluation needs of a community.

Evaluation goal: The desired knowledge about a system, its performance or usability.

Evaluation method: A series of steps or actions taken to accomplish an evaluation.

Evaluation model: An abstract, general or theoretical representation of an evaluation approach.

Evaluation parameter: A measurement or value on which something else in the evaluation process depends.

Evaluator: The person, group or organization that performs the evaluation.

Interactivity: The ability of complex systems to detect and react to human behavior.

Metric: A unit of measure for some property of a system.

Precision: The fraction of relevant documents in the set of all documents returned by a search.

Quality: The desirability of properties or characteristics of a system or process. Quality is contextual.

Qualitative: A type of empirical knowledge. Qualitative data are unbiased and/or relative.

Quantitative: Quantitative properties can be meaningfully measured using numbers; properties which are not quantitative are called qualitative.

Recall: The fraction of relevant material that is returned by a search.

Response time: The time a system or functional unit takes to react to a given input.

Test bed: A collection of resources that can be used for evaluation purposes.

Use case: A technique for capturing the potential requirements of a system or software. Each use case provides one or more scenarios that convey how the system should interact with the end user or another system to achieve a specific goal.

User situation: A state of the user in which an information need emerges. A user situation is formed by the contextual conditions of the user, both in the environment and the domain.

User task: The mental and physical activity of the user to achieve a goal.

Validation: Testing to ensure the system (or component) conforms to requirements.

Verification: Testing to ensure proper functioning of the system (or component) in technical terms.

References

1. Fuhr, N., Hansen, P., Mabe, M., Micsik, A., Solvberg, I.: Digital Libraries: A Generic Classification and Evaluation Scheme. In: Lecture Notes In Computer Science, vol. 2163, pp. 187–199. Springer, Berlin (2001) Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4–9, 2001. Proceedings

2. DELOS: DELOS Workshop on Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics. Technical report (2002) http://www.sztaki.hu/conferences/deval
3. Agosti, M., Fuhr, N. (eds.) Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, http://dlib.ionio.gr/wp7/workshop2004_program.html (2004).
4. Cleverdon, C.: The Cranfield tests on index language devices. In: Sparck-Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–59 Morgan Kaufmann, San Francisco (1997)
5. Turpin, A., Hersh, W.: Why batch and user evaluations do not give the same results. In: Proceedings of the SIGIR01, pp. 225–231. ACM, New York (2001)
6. CCSDS: A reference model for an open archival information system. Document Number: ISO 14721:2003 (2003) http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf
7. IFLANET: Functional requirements for bibliographic records (1998) http://www.ifla.org/VII/s13/frbr/frbr.pdf
8. Levy, D., Marshall, C.: Going digital: a look at assumptions underlying digital libraries.. Commun. ACM **38**, 77–84 (1995)
9. Borgman, C.: What are digital libraries? Competing visions. Inf. Process. Manage. **35**, 277–243 (1999)
10. DELOS: Reference models for digital libraries: actors and roles - final report. Technical report (2003) http://www.dli2.nsf.gov/internationalprojects/working_group_reports/actors_final_report.html
11. Agosti, M., Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Schek, H.J., Schuldt, H.: A reference model for DLMSs interim report. In: Candela, L., Castelli, D. (eds.) Deliverable D1.4.2 – Reference Model for Digital Library Management Systems [Draft 1], DELOS, A Network of Excellence on Digital Libraries – IST-2002-2.3.1.12, Technology-enhanced Learning and Access to Cultural Heritage – http://146.48.87.122:8003/OLP/Repository/1.0/Disseminate/delos/2006_WP1_D142/content/pdf?version=1 [last visited 2006, October 2] (2006)
12. Goncalves, M.A., Fox, E., Kipp, N., Watson, L.: Streams, structures, spaces, scenarios, societies (5S): a formal model for digital libraries. ACM Trans. Inf. Syst. **22**, 270–312 (2004)
13. Baker, S., Lancaster, F.: The measurement and evaluation of library services. Information Resources Press, Arlington (1991)
14. Nicholson, S.: A conceptual framework for the holistic measurement and cumulative evaluation of library services. J. Doc. **60**, 164–182 (2004)
15. Kyrillidou, M., Lorie, M., Hobbs, B., Choudhury, G., Webster, D., Flores, N., Heath, F.: Emerging tools for evaluating DL services: conceptual adaptations of LIBQUAL+ and CAPM. J. Digit. Inf. **4**, (2003) http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Heath
16. Kyrillidou, M., Giersch, S.: Developing the digiqual protocol for digital library evaluation. In: JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, pp. 172–173. ACM, New York (2005)
17. Tsakonas, G., Kapidakis, S., Papatheodorou, C.: Evaluation of user interaction in digital libraries. In: Agosti, M., Fuhr, N. (eds.) Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, Padua, Italy (2004) http://dlib.ionio.gr/wp7/workshop2004_program.html
18. Tsakonas, G., Papatheodorou, C.: Analyzing and evaluating usefulness and usability in electronic information services. J. Inf. Sci. 32(5), 400–419 (2006)
19. Marchionini, G.: Evaluating digital libraries: a longitudinal and multifaceted view. Library Trends **49**, (2000)
20. DELOS: A Network of Excellence on Digital Libraries (2004) Technical Annex 1, pag. 6 - DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618)
21. ISO: ISO 9241-11. Ergonomic Requirements for Office Work With Visual Display Terminals (VDT). Part 11: Guidance in Usability. International Standards Organization, London (1997)
22. DeGroot, S.P., Knapp, A.E.: Applying the user-centered design (ucd) process to the development of a large bibliographic navigation tool: a partnership between librarian, researcher and developer. Scopus White Paper Series (2004) http://www.elsevier.com/librarians
23. Blandford, A., Buchanan, G., Jones, M.: Usability of digital libraries. Int. J. Digit. Libr. **4**, 69–70 (2004)
24. Dickstein, R., Mills, V.: Usability testing at the University of Arizona Library: how to let the users in on the design. Inf. Technol. Libr. **19**, (2000)
25. Hilbert, D.M., Redmiles, D.F.: Extracting usability information from user interface events. ACM Comput. Surv. **32**, 384–421 (2000)
26. Ivory, M., Hearst, M.: The state of the art in automating usability evaluation of user interfaces. ACM Comput. Surv. **33**, 470–516 (2001)
27. Ke, H.R., Kwakkelaar, R., Tai, Y.M., Chen, L.C.: Exploring behavior of E-journal users in science and technology: transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. Libr. Inf. Sci. Res. **24**, 265–291 (2002)
28. Sfakakis, M., Kapidakis, S.: User behavior tendencies on data collections in a digital library. In: Lecture Notes in Computer Science, vol. 2458, pp. 550—559. Springer, Berlin (2002) Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002, Rome, Italy, September 16–18, 2002. Proceedings
29. Jones, S., Cunningham, S.J., McNab, R., Boddie, S.: A transaction log analysis of a digital library. Int. J. Digit. Libr. **3**, 152–169 (2000)
30. Sutcliffe, A.G., Ennis, M., Hu, J.: Evaluating the effectiveness of visual user interfaces for information retrieval. Int. J. Human-Comput. Stud. **53**, 741–763 (2000)
31. Kengeri, R., Fox, E.A., Reddy, H.P., Harley, H.D., Seals, C.D.: Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD. Int. J. Digit. Libr. **2**, 157–169 (1999)
32. Clark, J.A.: A usability study of the Belgian American Research Collection: measuring the functionality of a digital library. OCLC Syst. Serv. Int. Digit. Libr. Perspect. **20**, 115–127 (2004)
33. Rousseau, G.K., Rogers, W., Mead, S.E., Sit, R.A., Rousseau, J.B.A.: Assessing the usability of on-line library systems. Behav. Inf. Technol. **17**, 274–281 (1998)
34. Thong, J.Y.L., Hong, W., Tam, K.: Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences? Int. J. Human-Comput. Stud. **57**, 215–242 (2002)
35. Hartson, R.H., Shivakumar, P., Perez-Quinones, M.A.: Usability inspection of digital libraries: a case study. Int. J. Digit. Libr. **4**, 108–123 (2004)
36. Glosiene, A., Manzuch, Z.: Usability of ICT-based Systems: State-of-the-art Review. Technical Report CALIMERA Deliverable 9 (2004)
37. Dicks, R.S.: Mis-usability: on the uses and misuses of usability testing. In: Proceedings of the 20th Annual International Conference on Computer Documentation, New York, pp. 26–30 ACM, New York (2002)

38. Doubleday, A., Springett, M.V., Sutcliffe, A.G., Ryan, M.: A comparison of usability techniques for evaluating design. In: Proceedings of the Conference on "Designing Interactive Systems: Processes, Practices, Methods and Techniques", New York, pp. 101–110 ACM, New York (1997)
39. Hill, L.L., Carver, L., Smith, T.R., Frew, J., Larsgaard, M., Dolin, R., Rae, M.A.: Alexandria digital library: user evaluation studies and system design. J. Am. Soc. Inf. Sci. **51**, 246–259 (2000)
40. Bishop, A.P., Merkel, C., Neumann, L.J., Star, S.L., Sandusky, R.J., Ignacio, E.: Digital libraries: situating use in changing information infrastructure. J. Am. Soc. Inf. Sci. **51**, 394–413 (2002)
41. Fox, E.A., Heath, L.S., Rao, D., Brueni, D.J., Nowell, L.T., Wake, W.C., Hix, D.: Users, user interfaces, and objects: envision, a digital library. J. Am. Soc. Inf. Sci. **44**, 480–491 (1993)
42. Buttenfield, B.: Digital libraries: philosophies, technical design considerations, and example scenarios. In: Usability Evaluation of Digital Libraries, pp. 39–59. Haworth Press, New York (1999)
43. Blandford, A., Stelmaszewska, H., Bryann-Kinns, N.: Use of multiple digital libraries: a case study. In: Proceedings of the first ACM/IEEE-CS joint conference on digital libraries, pp. 179—188. ACM Press (2001)
44. Castillo, J., Hartson, H.R., Hix, D.: Remote usability evaluation: can users report their own critical incidents. In: CHI 1998 Conference Summary on Human Factors in Computing Systems, New York, pp. 253–254 ACM, New York (1998)
45. Bates, M.: Where should the person stop and the information search interface start? Inf. Process. Manage. **26**, 575–591 (1990)
46. Cousins, S.B.: A task-oriented interface to a DL. In: Tauber, M.J. (ed.) Conference Companion on Human Factors in Computing Systems: Common Ground, New York, pp. 103–104. ACM, New York (1996)
47. Pejtersen, A., Fidel, R.: A framework for work centered evaluation and design: a case study of IR on the web. Technical report (1998)
48. Rasmussen, J., Pejtersen, A., Goodstein, L.: Cognitive Systems Engineering. Whiley, New York (1994)
49. Borgman, C.: From Gutenberg to the global information infrastructure: Access to information in the networked World. MIT, Cambridge (2000)
50. Tenopir, C.: Use and users of electronic library resources: an overview and analysis of recent research studies. Council on Library and Information Resources, Washington (2003)
51. Banwell, L., Ray, K., Coulson, G., Urquhart, C., Lonsdale, R., Armstrong, C., Thomas, R., Spink, S., Yeoman, A., Fenton, R., Rowley, J.: The JISC user behaviour monitoring and evaluation framework. J. Document. **60**, 302–320 (2004)
52. Payette, S.D., Rieger, O.Y.: Supporting scholarly inquiry: incorporating users in the design of the digital library. J. Acad. Libr. **24**, 121–129 (1998)
53. Yu, L., Apps, A.: Studying E-journal user behavior using log files: the experience of superjournal. Libr. Inf. Sci. Res. **22**, 311–338 (2000)
54. Wilson, T.D.: Information behaviour: an interdisciplinary perspective. Inf. Process. Manage. **33**, 551–572 (1997)
55. Kuhlthau, C.C.: Inside the search process: information seeking from the user's perspective. J. Am. Soc. Inf. Sci. **42**, 361–371 (1991)
56. Ellis, D., Haugan, M.: Modeling the information seeking patterns of engineers and research scientists in an industrial environment. J. Document. **53**, 384–403 (1997)
57. Foster, A., Ford, N.: Serendipity and information seeking: an empirical study. J. Document. **59**, 321–340 (2003)
58. Belkin, N.J., Oddy, R.N., Brooks, H.M.: ASK for information retrieval: part 1 Background and theory. J. Document. **38**, 61–71 (1982)
59. Wilson, T.D.: Exploring models of information behaviour: the uncertainty project. Inf. Process. Manage. 893–849 (1999)
60. Wood, R.: The Conspectus: a collection analysis and development success. Libr. Acquis. Pract. Theory **20**, 429–453 (1996)
61. Saracevic, T.: Evaluation of digital libraries: an overview. In: Agosti, M., Fuhr, N. (eds.) Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, Padua, Italy (2004) http://dlib.ionio.gr/wp7/workshop2004_program.html
62. Agosti, M., Ferro, N., Frommholz, I., Thiel, U.: Annotations in digital libraries and collaboratories: facets, models and usage. In: Lecture Notes In Computer Science, vol. 3232, pp. 244—255 Springer, Berlin (2004) Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004. Proceedings
63. Agosti, M., Albrechtsen, H., Ferro, N., Frommholz, I., Hansen, P., Orio, N., Panizzi, E., Pejtersen, A.M., Thiel, U.: Dilas: a digital library annotation service. In: International Workshop on Annotation for Collaboration. Methods, Tools, and Practices, pp. 91–101 (2005)
64. Hansen, P., Jarvelin, K.: The information seeking and retrieval process at the swedish patent and registration office: moving from lab-based to real life work task environment. In: Agosti, M., Fuhr, N. (eds.) Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval, Athens, Greece (2000) 43–53 http://www.sics.se/~preben/papers/SIGIR20000-WS.pdf
65. Covey, D.: Usage and usability assessment: library practices and concerns, vol. 105. CLIR (2002) http://www.clir.org/pubs/reports/pub105/pub105.pdf
66. Borgman, C.L.: Social Aspects of Digital Libraries-Final Report to the NSF. Technical report (1996) http://is.gseis.ucla.edu/research/dl/UCLA_DL_Report.html
67. Kovács, L., Micsik, A.: The evaluation computer: a model for structuring evaluation activities. In: Agosti, M., Fuhr, N. (eds.) DELOS Workshop on the Evaluation of Digital Libraries, Padova, Italy (2004) http://dlib.ionio.gr/wp7/workshop2004_program.html
68. Banwell, L., Coulson, G.: Users and user study methodology: the JUBILEE project. Inf. Res. **9**, 47–56 (2004) http://informationr.net/ir/9-2/paper167.html
69. Jeng, J.: What is usability in the context of the digital library and how can it be measured? Inf. Technol. Libr. **24**, 47–56 (2005)
70. Xie, H.: Evaluation of digital libraries: criteria and problems from users perspectives. Libr. Inf. Res. **28**, 433–452 (2006)
71. Klas, C., Albrechtsen, H.N.F., Hansen, P., Kapidakis, S., Kovacs, L., Kriewel, S., Micsik, A., Papatheodorou, C., Tsakonas, G., Jacob, E.: A logging scheme for comparative digital library evaluation. In: Lecture Notes In Computer Science, vol. 4172, pp. 267—278 Springer, Berlin (2006) Research and Advanced Technology for Digital Libraries: 10th European Conference, ECDL 2006, Alicante, Spain, September 17—22, 2006. Proceedings