DEVELOPING A HOLISTIC MODEL FOR DIGITAL LIBRARY EVALUATION

by

YING ZHANG

A Dissertation Submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Communication, Information, and Library Studies

written under the direction of

Dr. Tefko Saracevic

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2007

ABSTRACT OF THE DISSERTATION

Developing a Holistic Framework for Digital Library Evaluation

by YING ZHANG

Dissertation Director:

Dr. Tefko Saracevic

The objective of the research is to develop a holistic model for digital library (DL) evaluation. To develop such a model, a three-stage research approach was applied: exploration, confirmation, and verification. During the exploration stage, a literature review was conducted, and then an interview along with card sorting technique was employed to collect perceptions from DL experts with emphasis on determining what criteria should be used in DL evaluation. Then, the criteria identified from the exploration were used for developing an online survey during the confirmation stage. Heterogeneous DL stakeholders were asked to rate the importance of each criterion to DL evaluation. The holistic model was constructed by utilizing descriptive and inference statistical techniques. Its holistic nature was ensured through: (1) incorporation of various DL stakeholders' perspectives in light of Marchionini's multifaceted evaluation approach, and (2) inclusion of all digital library levels suggested by Saracevic's stratified information retrieval model. Eventually, in the verification stage, selected criteria from the model were tested in real DL use setting.

Some significant findings include: (1) consistently perceived important criteria for DL evaluation. DL stakeholders care more about premise (e.g., *accessibility* and

sustainability of a DL), process (e.g., *ease of use, technology reliability,* and *service responsiveness*), and direct performance (e.g., *usefulness of information, successfulness* and *efficiency of task completion*), whereas less concerned about indirect factors (e.g., *personalization, behavior change, service courtesy,* and *extended social effects*); (2) inter-group divergence in importance perception for some evaluation criteria. The divergence primarily exists between the user and other DL stakeholder groups; (3) some promising criteria (e.g., *comprehensiveness of collection, integrity of information, integration of service to information seeking path, collaboration/sharing*) augment the existing DL evaluations whereby important criteria have essentially been covered; and (4) most importantly, the core dissertation objective is fulfilled, that is the construction of the holistic evaluation model, in which heterogeneous stakeholders' perspectives at all DL levels are presented.

The proposed model fills a lacuna in the DL domain, that is the lack of a comprehensive and flexible framework to guide and benchmark evaluations and the uncertainty about what divergence exists among heterogeneous DL stakeholder groups.

ACKNOWLEDGEMENT

It has been a long, hard but cherished journey towards the Ph.D. degree. During the past six years, numerous friends have helped me, both physically and mentally. I am indebted to them for their unconditional kindness and support.

Firstly, I want to thank my adviser, Dr.Tefko Saracevic, whose outstanding advisorship has been inspiring me throughout the journey. Further, I want to thank my committee members, Dr. Heting Chu, Dr. Michael Lesk, Dr. Claire McInerney, as well as Dr. Daniel O'Connor, for their insightful comments and constructive criticism on my dissertation proposal and drafts.

Secondly, I would like to thank my friends within and without the School of Communication, Information and Library Studies, for their dedication in assisting my research and daily life. In addition, I thank all participants in my studies for their time and effort.

Lastly, I would like to say "thank you" to my son, for his understanding and support, as well as to my parents and my uncle and aunt, for their love and caring throughout the dissertation process.

TABLE of CONTENTS

List of Tables

List of Illustrations

Chapter 1 INTRODUCTION

1.1 Digital Libraries

Nearly 60 years ago, an American scientist, Vannevar Bush (1945), drew a vivid picture showing how the machine, Memex, could be used to store, organize and retrieve information as well as to represent the epistemological association among human minds. Similarly, Licklider (1964) had an inspiring dream about how advanced information technologies could dramatically change the ways in which a traditional library serves it users. These two can be seen as scientific predictions foretelling the birth of the digital library (DL).

The World Wide Web, along with advanced computation technologies, catalyses DL research and practices. The early ’90s saw the growth of DL activities. Since then, a number of researchers and professionals from different disciplines (e.g., information technology, computer science, information science and librarianship) have been attracted to this area. “Digital libraries have a short yet turbulent and explosive history” (Saracevic, 2000 p.350). The past decade saw an exponential increase in the number of ongoing and completed DL projects. In addition to the pace and scale of DL projects, the activities and entities around DL development and research are astonishing as well. There have been a number of journals (e.g., *D-Lib Magazine, International Journal of Digital Libraries*, *Journal of Digital Information*), proceedings (e.g., *ACM Joint Conference on Digital Libraries*, *European Conference on Research and Advanced Technology for Digital Libraries*, *International Conference on Asian Digital Libraries*), and organizations (e.g., *Digital Library Federation*) that specifically target this topic. Moreover, some prestigious journals also have devoted special issues to this subject (e.g.,

Journal of American Society for Information Science 1993, 2000; *Information Processing and Management*, 1999; *Library Trends*, 2000; *Communication of ACM,* 1995, 1998. *IEEE Computer,* 1995).

However, behind this fast-growing scene are some weaknesses that might hinder the progress of DL innovation. One of the remarkable weaknesses is reflected in the two competing visions of what is a DL, which are pinpointed by Borgman in 1999. One vision, with the research domain (e.g., Fox et al., 1993; Lesk, 1997) as the representative, sees DL as a digital collection combining computing, storage and communication technologies with distributed physical contents. On the other hand, another vision can be mainly found from the professional domain. For instance, Waters in his 1998/1999 Digital Library Federation Annual Report, defines the DL as the institutional extension of traditional libraries in digital environments. In Borgman's view, both visions are "problematic" because they constrain the boundaries between digital collections and institutions.

Unsurprisingly, the divergent visions of the definition of the DL cannot but yield to different foci in DL development. The view of the DL as a collection highlights digital information and related information technologies in terms of preservation, storage, representation, and retrieving. In comparison, the organization vision puts heavy weight on service through which one group of people satisfies the needs of another group in a socialized setting. Echoing Borgman's viewpoint (1999), Saracevic (2000) sees the two visions situated at "the two ends of a spectrum," and they would be better if they could meet at the middle.

The biased visions might impede the effective development of a DL, because any of the elements (i.e., collection, technology, service, people, and the contexts associated with the elements above) are crucial to an operational DL. DLs are catalyzed by advanced communication and computer technologies. DL cannot exist without technology. Compared with conventional information retrieval (IR) systems and libraries, DLs usually attract more diverse groups of people in addition to target users, IT scientists, and LIS communities, such as archivists, publishers, funders, administrators, and so forth. Furthermore, considering that a DL usually involves an ever-massive amount of intellectual and monetary resources, it is essential to have a well-developed infrastructure for facilitating functions and services within and among DLs.

As a matter of fact, whereas a DL may share common characteristics with traditional IR systems (e.g., information organization, representation, retrieval, the interaction between a system and its users) and libraries (e.g., collections management, services), it has unique features as well. They are: (1) highly dynamic and ephemeral in technical, collection and information needs (Fox & Urs, 2002), (2) highly heterogeneous along digital format, collection coverage, user and system dimensions, (3) tightly virtual or non-virtual collaboration among different groups of people, including knowledge creators, publishers, distributors, information specialists, librarians, and users. (O'Day & Nardi, 2003), and (4) environmental (i.e., technical, economic, legal, institutional, social) dependency.

1.2 Digital Library Evaluation

Accordingly, on the one hand, DL evaluation may borrow approaches and criteria from those being used in the evaluations of traditional IR system and libraries. On the

other hand, it is essential to develop a specific DL evaluation framework. Furthermore, with an enormous consumption of technical, financial and personnel resources, each DL project ought to be evaluated to secure the outcome of its development.

Unfortunately, compared with the growing number of DL projects, the overall quality of DLs is insufficiently studied and reported (Chowdhury & Chowdhury, 2003; Saracevic, 2000; Xie, 2006). The problem is presumably associated with fewer evaluation studies that assess how well (e.g., effective, efficient, usable, useful, etc.) these DLs have been implemented to help different groups of users locate desired resources in particular contexts. "Evaluation is more conspicuous by its absence (or just minimal presence) in the vast majority of published work on digital libraries… So far, evaluation has not kept pace with efforts in digital libraries" (Saracevic 2000 p.351). Furthermore, the evaluation approaches and criteria vary among the existing DL evaluation studies. The majority of the studies adopt traditional IR evaluation approaches at a restricted level (either at the system or the user level) while employing traditional criteria, such as precision/recall, search time, error rate, etc. Very few address the effects of a DL at higher levels (e.g., social, legal, cultural) in terms of how well a DL fits into or even improves people's daily work/life. Furthermore, there are few metrics devised specifically for DL settings.

Consequently, the existing evaluation products fail to uncover the state of the art of DL innovation. Additionally, it is difficult to compare and contrast among different DLs, considering variations in evaluation metrics and methods. Having acknowledged the lacuna, a number of professionals and researchers have been seeking a valid DL evaluation framework, suggesting what should be evaluated, how a DL should be evaluated, and who should evaluate it. In 1998 (July/Aug), *D-Lib Magazine* published a

report by the Computer Science & Telecommunication Board, National Research Council, within which the following conclusion is clearly stated and heuristic to DL evaluation: "Reaching a consensus on even a minimum common denominator set of new statistics and performance measures would be a big step forward…" Borgman (2002b) straightforwardly commented: "the digital library community needs benchmarks for comparison between systems and services… We also need a set of metrics for comparing digital libraries."(p.10).

Chapter 2 PREVIOUS STUDIES

The literature review reveals the weaknesses in DL evaluation research and practices. To investigate the status of DL evaluation, a search for relevant papers was performed in key databases and systems in the domain of Library and Information Science (LIS), including Library & Information Science Abstracts, Information Science Abstracts, Library Literature & Information Science, ACM Digital Libraries, and IEEEXplore. Additionally, the Web of Science (WoS) was examined to expand the search scope from the dominant body of DL research and development to plausible DL application domains (e.g., education, health). It should be noted that WoS is a multidisciplinary database that indexes research from leading journals in each discipline. Therefore, the WoS search should be representative rather than comprehensive. Chapter 4, the Methodology Chapter, will provide a detailed rationale for the literature review approach.

The literature review was focused on DL evaluation frameworks (i.e., what is the standpoint evaluators take), criteria (i.e., what to be evaluated), and/or methodologies (i.e., how evaluations have been conducted).

2.1 Frameworks

To have a framework within which an evaluation is implemented is essential to improve research outcomes. There are four main evaluation frameworks identified from the literature: user-oriented, system-oriented, systematic, and longitudinal. The following four sections summarize some representative researches and pros/cons for each framework. It should be noted that the categorization of the research aims to highlight the

primary evaluation theme in a given study. As such, the categorical assignments may not be mutually exclusive.

2.1.1 User-oriented Evaluation

Marchionini et al. (2003) argue that DL evaluation "must be rooted" in information needs, characteristics, and contexts of potential users, because DLs are built to "serve communities of people". Van House et al. (1996) recommend that a truly useful DL design has to be aware of all user related factors, such as a larger context for information needs, purposes for using DLs, individual users' specific tasks, etc.

This approach has been most frequently used in examining how interfaces are designed to meet users' needs, behaviors and preferences, and how users interact, use, perceive and are satisfied with a given aspect (e.g., collection) of a DL. Target users are usually gatekeepers (participants/interview participants) of evaluations (Abbas, 2002; Bishop, 1999; Borgman et al., 2001; Carter et al., 2000; Fox et al., 1993; Hill et al., 2000; Jones et al., 2000; Marchionini, 1998; Mead, 1995; Park, 2000; Peng et. al., 2004; Sumner, 2001; Thong, 2002; Wildemuth et. al., 2003). In general, the research findings show that the performance of DLs is associated with system features/functions as well as users' characteristics (e.g., information needs type, information skill, subject knowledge background, cognitive mode). Thong et al. (2002) collected data from 397 users of an award-winning DL. They found that whereas perceived usefulness and ease of use determine user acceptance of a DL, individual differences (i.e., computer self-efficacy, computer experience, domain knowledge) affect perceived ease of use.

The user-oriented evaluation approach is enlightening in terms of having highlighted the direct purpose of DL innovation, which is to help users find their desired

information for given purposes. The results of a user-oriented evaluation are more likely to be useful for DL system improvement in terms of meeting target users' needs and fitting into their background. Nevertheless, while focusing on information searching related characteristics, a user-oriented evaluation approach lacks a way to address all user related factors, especially environmental factors that may shape a given user's information needs, information seeking behaviors, and information use. Additionally, users usually do not have the entire picture of a system or a collection, such as missing records or duplicate records in a collection. Consequently, it is impossible for users to evaluate a DL system as a whole.

2.1.2 System-oriented Evaluation

Compared with the user-oriented approach, the system-oriented evaluation framework has been adopted less often in DL evaluations. The small number of studies is primarily from the computer domain, where the research focus is on technological innovation rather than use. These studies aim to examine how well advanced technology can be used for digital information (mostly audio and video information) representation and retrieval. One may see some representative cases in light of this approach from the TREC Video Track (Hauptmann, 2001; Ma, 2001; Smeaton, 2001; Hidaka, 2001). Other instances of system-oriented evaluation can be found at the ACM Multimedia conference (Rui, 2000; Zhang, 1995) and IEEE International Workshop on Database and Expert Systems Applications **(**Hee, 1999**)**. The traditional pair of precision/recall and search time is the most frequently used metric for this school of evaluation research, where target users are rarely involved.

The system-centered approach assumes that as long as a DL is implemented with high standards, it will surely satisfy its users. Nevertheless, the assumption is not justified by the research findings from the preceding section of the user-centered approach, where user related factors do influence the outcome of DL uses (Baldonado, 2000; Borgman et al., 2001; Carter et al., 2000; Hill et al., 2000; Jones et al., 2000; Marchionini, 1998; Sumner, 2001; Thong, 2002).

2.1.3 Longitudinal Evaluation

Whereas the system-oriented and the user-oriented approaches highlight the influences of a given aspect (or several aspects) on the output/outcomes of a DL at a particular information searching stage, a longitudinal approach pays extra attention to DL's temporal effects. This approach assumes that a DL usually exerts its impact on its user community over years. Some effects might not be observable at the beginning of the DL application. Furthermore, rapid technological innovations often push system changes of a DL, and the system changes may then yield to the changes of users' information needs, behaviors, and use. Accordingly, "the evaluation plan has to be a roadmap that would guide decision-making over the years…" (Marchionini, 2000, p.313). This depiction is based on the evaluation experiences from the Perseus Digital Library (PDL), where substantial changes have been examined over 10 years. According to Marchionini (2000), the longitudinal evaluation is essential, because the ultimate goal of a DL is to change users' way of living and the larger social milieu, which might not be achieved in the short run. For instance, the educational effect of Perseus was not recognized until after a few years of use.

The Alexander Digital Earth ProtoType (ADEPT) at the University of California at Los Angels (UCLA) and the University of California at San Barbara (UCSB) is another case where a five-year long, formative evaluation study was conducted (Borgman et al, 2000). Whereas the results from the initial stage (2000-2001) showed that ADEPT had impacts on the teaching and learning patterns at the undergraduate level, the 2nd stage (2001-2004) was expected to see the changes in students' scientific thinking and learning. Similarly, in a report from the Computer Science & Telecommunication Board of the National Research Council (1998), the need for "early and often evaluation" involving representative users was highlighted, based on the review of a number of ongoing information system projects.

The longitudinal evaluation framework is promising in terms of being able to address long-term environmental impacts of DLs in addition to a narrower sense of DL output and outcome within the context of facilitating individual users' task implementations. Moreover, the stage-by-stage formative evaluation findings are more likely to provide timely feedback to DL development. In particular, it is suitable for assessing a large scale DL project that involves years of efforts and tremendous amounts of human and financial resources.

2.1.4 Systematic Evaluation

Whereas a longitudinal approach highlights temporal effects of DLs at various design and use stages, a systematic evaluation focuses more on "spatial" effects at various DL levels as well as from different DL stakeholders' viewpoints.

Based on his stratified IR Interactive model proposed in 1996, Saracevic (2000) suggests a conceptual framework that outlines the constructs and contexts of DL

development. According to the author, DLs can be evaluated in different contexts (e.g., content, engineering, interfaces, user, environmental, etc.) targeting various constructs (e.g., collection, connection, organization, representation, preservation, access, distribution, interaction, search, service, assistance, use). In the author's view, a successful DL evaluation should take all these constructs and contexts into account, although an individual study may take any slice from them. The DL design layers from Bates's Cascade Model (2002), from network, computer system, information content, interface, to user search activities, user understanding and motivation, are similar to Saracevic's stratified context and constructs. While asserting the interrelation among these layers, the author argues that "all layers … should be simultaneously designed with knowledge of what is going forward in the other layers" and "digital libraries cannot be fully effective as information sources for users until the entire design process is done in a manner that involves genuine conceptual and practical coordination among the people working on the system layers." (Bates, 2002, p397) Although the author claims that the model is developed for DL design purpose, the various layers can be treated as systematic DL evaluation objects. The only weakness of the model is the exclusion of environmental (e.g., organizational, social, culture) layers. Likewise, Fuhr et al. (2001) proposed a holistic evaluation scheme for DL evaluation. In addition to the same weakness as Bates' cascading model in terms of excluding larger impacts of DLs other than influences on individual users and user groups, this scheme neglects another crucial DL level –interface. Consequently, from a systematic standpoint, the scheme, comprised of data/collection, system/technology, users and usage, cannot be regarded a holistic model.

In addition to the emphasis on the interplay among different DL levels and aspects, the consideration of contextual effects is another key feature of a systematic approach, distinguished from the user-centered and system-centered approaches. Thong et al.'s survey results (2002) suggest that in addition to individual users' characteristics, organizational context (relevance, system accessibility, and system visibility) is another independent variable, which proved to impact perceived ease of use and usefulness of the DL. Similarly, Adams & Blandford's (2001) focus group discussions and in-depth interviews in clinical settings reveal that the perceived impacts of DL innovation are associated with organizational, social and political structures. Also, organizational hierarchies impede the use of DLs. While not particularly targeting DL evaluation, Wallace (2001) argues that failure in understanding the context of system development and use may yield to evaluation activities "inappropriate, ineffective, or even harmful". Grounded on rich and in-depth arguments and evidence about the DL as a socio-technical system, Bishop et al. (2003) advocates for "technically informed social analysis" for DL evaluation.

Furthermore, several studies suggest that a convincing DL evaluation should not only consider the output/outcome of a given DL aspect, but also incorporate the input factor. For instance, Kantor and Saracevic (1999) devised a measure for assessing various library services in the digital age. The service value is a function of *time spent* against *perceived benefit*. Although the measurement might be too simplified to embrace all value-associated factors, it tends to be a more convincing measure for a service assessment. Similarly, Bekele (2002) argues for the necessity of using *cost-effectiveness* as a measure for evaluating digital collection and incorporation through a case study of

the OSSREA digital library. The measure and argument is in line with ARL (Association of Research Libraries)'s e-metrics and EU EQUINOX projects where *cost* is proposed as one of the measures for assessing research library performance in the digital age.

In addition to these systematic perspectives emphasizing the inclusion of multidimensional DL aspects as evaluants, several other scholars highlight the variety of DL stakeholders and the need to include their diverse perspectives in evaluations. Nicholson, S. (2004) proposed a conceptual framework for the holistic measurement and cumulative evaluation of library services. According to the author**,** both system and user views of evaluation are essential to the library service evaluation. The same criteria should and could be judged in different ways by different participants (e.g., users, library personnel, and decision-makers). This viewpoint echoes Marchionini (2000)'s multifaceted DL evaluation framework. The essence of the latter resides in systematic data collection on and integration of different viewpoints, using different approaches and from different dimensions.

Despite these sound advocates for a systematic approach, so far, very few DL projects have been found to have their evaluation work done in light of the framework. The majority of DLs were evaluated merely at one or two levels. Primary criteria used were arbitrarily determined by DL researchers and/or developers. Among the handful projects with a somewhat systematic standpoint, the multifaceted evaluation for the Perseus Digital Library (PDL) is worth mentioning, where the evaluation has been done at four levels: users (learners & educators), technical systems, content, and educational (Machionini, 2000). The evaluation provides a comparative holistic picture of how well the PDL has been implemented.

2.2 Criteria/Measures

The systematic standpoint demands that DL evaluation should have a holistic set of evaluation metrics, which can be isomorphically matched to different DL levels and aspects. Having argued for the necessity of assessing each DL project, Saracevic (2000) goes further, suggesting that each evaluation has to be conducted with essential decisions on construct, context, criteria, measures, and methodology. It would be better to have clearly pre-determined elements of criteria, measures, methodologies as well as of the larger "view" of construct and context of evaluation. Alternatively and constructively, as pinpointed by the author, DL evaluation might borrow criteria and measures from the domains of traditional library, information retrieval, and human-computer interaction. By looking at the list of these criteria (see Table 2.1) one may see that the majority of them have been employed as indicators in current DL evaluations (see Table 2.2 through Table 2.7 in the following sub-sections). However, some criteria (e.g., intellectual protection and standard accordance) have drawn less attention, while others (e.g., usability and relevance) have been used more frequently.

Table 2.1: Suggested criteria from other domains for DL evaluation (Saracevic, 2000)

Traditional library	Traditional IR	Traditional user interface
• Collection: purpose, subject, scope, coverage, authority, currency, audience, cost, format, treatment, preservation and persistence • Information: accuracy, appropriateness, links, representation, uniqueness, compatibility, presentation, timeliness and ownership • Use: accessibility, availability, searchability, and usability • Standards	• Relevance (P/R, etc.) • Satisfaction • Index, search and output features	• Usability, functionality, effort • Task appropriateness, failures • Connectivity, reliability • Design features • Navigation, browsing • Services, help

In addition to the unbalanced use of conventional criteria in existing evaluations, DL specific metrics are inadequate as well. For instance, very little research has devised

effective measures to examine how well a digital object is created as a representation of its original artifact, how well a DL system meets the needs of various user group(s) and supports the social interaction in a larger context, and how well heterogeneous digital information and systems are seamlessly integrated and interoperated. One may further see the problematic situation from the following sub-sections regarding evaluation criteria and measures for different levels of DL evaluation (i.e., information/collection, hardware/software, interface, service, user, context).

2.2.1 Content Level Evaluation Criteria/Measures

With the proliferation of electronic sources in the digital age, the quality of information and collection is likely to be one of the key factors influencing the outcome of a DL. How well digital objects are selected and created, how well digital information is organized and represented, to what extent a DL collection meets target users' information needs, and how well digital information/collection is associated with other content can be the main objectives at this level. Nevertheless, it seems that this body of evaluation is one of the weakest parts in existing DL evaluations. Few studies report their DL evaluation at this level, while some studies (e.g., Xie, 2006) have suggested the significance of digital content evaluation. Table 2.2 lists the criteria that have been used in or proposed for (in italic font) digital content evaluation.

Table 2.2: Existing Criteria for DL Evaluation – Content

Digital object

- Fidelity [Jones et al., 1999; Kenney et al., 1998]

Metadata

- Adequacy [Huxley, 2002]
- Diversity [Borgman et al., 2001; Fuhr et al., 2001]
- Extensibility [Borgman et al., 2001]
- *Standardization [Kwak et al., 2002]*
- Suitability to original artifact [Goodrum, 2001; Jones & Paynter, 2002]

Information

- Accessibility [Adams & Blandford, 2001; Bishop, 1998; Jones et al., 1999; Wilson et al., 2002]
- Accuracy [Bergmark et al., 2002; Jones et al., 1999; Kwak et al., 2002; Machionini et al., 2003; Zhang et al., 1995]
- *Appropriateness for target/potential audience* [Borgman et al., 2001; Ding et al., 1999]
- Authority [Budhu & Coleman, 2002]
- Clarity [Hill et al., 1997; Huxley, 2002, Kengeri et al., 1999]
- Conciseness [Rittman et al., 2004]
- Connection to other appropriate resources [Kwak et al., 2002]
- Copyright protection /ownership [Besek, 2003]
- Cost [Budhu & Coleman, 2002; Choudhury et al., 2002; Brophy et al., 2000; Shim, 2000; Fuhr et. al., 2001]
- Ease of understand and learn [Borgman et al., 2000, Khoo et al., 2002, Zhang et al., 2004]
- Informativeness [Huxley, 2002; Zhang et al., 2004]
- Novelty [Larsen, 2000]
- Potential distraction [Sumner et al., 2003]
- Readability [Kengeri et al., 1999]
- Scalability [Kengeri et al., 1999; Kenney et al., 1998; Larsen, 2000]
- *Supportiveness of* human-computer [Budhu & Coleman, 2002] *& social/group interaction* [Borgman & Gilliland-Swetland, 2000]
- Timeliness (freshness) [Bekele, 2002; Kwak et al., 2002]
- Usefulness (Zhang et al., 2004)

Collection

- *Comprehensiveness (in subject, time, language etc.) [Jones et al., 1999; Kwak et al., 2002; Kengeri et al., 1999]*
- Coverage diversity (in subject, time, format, etc.) [Blixrud, 2002; Budhu & Coleman, 2002; Meyyappan et al., 2000]
- Cost-effectiveness [Bekele, 2002]
- *Growth rate* [Kwak et al., 2002]
- Size [Blixrud, 2002; Franklin, 2002]

Essentially, these criteria are employed to assess four types of digital content: digital object, metadata, information, and collection. Among these four, the evaluation of digital objects seems to be the only unique topic which can be found in DL contexts, whereas the other three entities have been evaluated more or less with conventional criteria (e.g., *accuracy, authority, clarity, cost, ease of understanding, informativeness,*

readability, timeliness, usefulness). Kenney et al. (1998) reports their digitization work at the Library of Congress and at Cornell University. The quality of digitized objects was examined in terms of how well the digital version captures the *essence, detail*, and *structure* of the original artifacts. From a technical perspective, these indicators are strong in assessing the quality of digitization work.

Specifically for DL content evaluation in broader views, there are some other metrics worthnoting, such as *scalability for different user communities* (Kengeri et al., 1999; Kenney et al., 1998; Larsen, 2000)**,** *potential distraction* in educational setting (Sumner et al., 2003), as well as *supportiveness for social /group interaction* (Borgman & Gilliland-Swetland; 2000). These metrics tackle some crucial issues in DL innovation. For instance, as noted above, compared with conventional IR systems and libraries, a DL usually has diverse user communities with various backgrounds and changing needs. As such, it is essential to provide scalable contents including different organization and presentation options to each of the target user groups. Additionally, DLs are created not only in an information seeking context, but also with a promise of integrating itself into the whole society. Hence, one has to evaluate a DL from social (*supportiveness of social interaction*), and cultural (*potential distraction in class*) perspectives.

In spite of these sound metrics, overall, there is still something missing in DL content evaluations, such as examining how well a DL meets users' information needs and how well digital meta-information schemas reflect the ontology of the discipline(s) that a given DL covers.

2.2.2 Technology Level Evaluation Criteria/Measures

So far, digital technology evaluation has been focused on two aspects, namely hardware and software. The former is geared to examining the extent to which up-to-date computers and networks may improve the quality of digital information presentation and optimize information processing and transmission. The latter highlights primarily conventional relevance-based effectiveness measures (see Table 2.3). It should be noted that some researchers (Hee et al., 1999; Salampasis et al., 2002) have modified conventional effectiveness measures to fit into digital and hypermediated circumstances. For instance, when evaluating a video DL, Hee et al. (1999) adjusted *the P/R pair* as "user-defined relevant scenes n-retrieved relevant scenes over retrieved scenes" and "user-defined relevant scenes n-retrieved relevant scenes over user defined relevant scenes" respectively. The new adapted measures are nicely accommodated to the video search setting, whereby relevance is more likely judged upon clips and scenes rather that the whole video tape.

In addition to these effectiveness measures, *reliability, cost,* and *response time* are used for both hardware and software evaluations. However, considering the existence of diverse DLs and the necessity of integrating these systems, it is disappointing to see no evaluations on *interoperability/compatibility* among different DL systems, although the metrics are included in Kwak et al.'s evaluation model (2000) for university libraries in the digital age. Kwat et al.'s model is a product of their two-phased research (i.e., existing work survey + thrice-run Delphi) grounded on opinions from DL experts and researchers. Similarly, *auxiliary functionality* (e.g., privacy protection, firework) receives little attention.

Table 2.3: Existing Criteria for DL Evaluation – Technology

Hardware

- Accessibility [Bishop, 1998; Kwak et al., 2002; Meyyappan et al. 2000; Wilson et al., 2002]
- Appropriateness for digital information [Kwak et al., 2002]
- Comfort for use [Wilson et al., 2002]
- Cost [Thebridge et al., 2002]
- Display quality [Wilson et al., 2002]
- Efficiency (number of node utilization [Xi et al., 2002]; response time [Fuhr et al., 2002; Kengeri et al., 1999; Larsen, 2000]; network related response time [Kapidakis et al., 1998; Kwak et al., 2002])
- Robustness for digital information [Marchall & Ruotolo, 2002; Wilson et al., 2002]

Software

- Accessibility [Bishop, 1998; Kwak et al., 2002; Meyyappan et al. 2000; Wilson et al., 2002]
- *Auxiliary functionality* (e.g. privacy protection, firework [Kwak et al., 2002])
- Complexity in query support and response [Larsen, 2000; Meyyappan et al. 2000]
- Cost [Thebridge et al., 2002]
- Efficiency (e.g. number of node utilization [Xi et al., 2002]; response time [Fuhr et al., 2002; Kengeri et al., 1999; Larsen, 2000]; network related response time [Kapidakis et al., 1998; Kwak et al., 2002])
- *Interoperability / compatibility (among different IR and DL system [Kwak et al., 2002]*
- Relevance-based effectiveness (e.g. number of relevant document retrieved [Jones & Lam-Adesina, 2002; Khoo et al., 1998], P/R [Bosman et al., 1998; Jones & Lam-Adesina, 2002; Hee 1999; Khoo et al., 1998; Sanderson & Crestani, 1998], Actual vs. perceived P/R [Larsen, 2000]; accuracy of surrogate extraction [Rui et al., 2000]; relative distance relevance [Salampasis et al., 2002])
- Reliability [Papadakis et al., 2002]; stability of system [Champeny et al., 2004]

One might find that some system criteria listed in Table 2.3 (e.g., *display quality, robustness for digital information*) tends to evaluate electronic and communication devices rather than DLs. However, considering the high dependency of DLs on advanced technologies, it is likely that DL effects could be largely influenced by these factors.

2.2.3 Interface Level Evaluation Criteria/Measures

Interface is one of the most significant levels in a DL, because it is the surface where a system and its users meet. A number of studies (e.g., Abbas, 2000; Baldonado, 2000; Park, 2000) have been conducted at this level. In general, there are three primary evaluation objectives: (1) how effective and efficient a DL is in terms of helping users find needed information; (2) how well the interface fits users' knowledge background and information seeking needs/behavior; and (3) to what extent the interface is in accordance with interface design principles.

Table 2.4: Existing Criteria for DL Evaluation – Interface / Interaction

- Aesthetics (e.g., attractiveness, simplicity) [Budhu & Coleman, 2002; Hill et al., 1997; Thong et al., 2002; Wesson, 2002; Zhang, 2004]
- Appropriateness to target users [Dillon, 1999; Zhang, 2004]
- Availability of additional assistance (e.g. help, example search, feedback) [Huxley, 2002; Kengeri et al., 1999]
- Consistency [Salampasis et al., 2002; Wesson, 2002; Zhang, 2004]
- Ease of use [Champeny et al., 2004; Huxley, 2002; Jeng, 2005; Khoo et al., 2002]; navigation [Hill et al., 1997, Huxley, 2002; Papadakis et al., 2002]; understanding [Khoo et al., 2002])
- Effort [Jeng, 2005; Larsen, 2000; Zhang et al., 2004]
- Error detection/ handling; Error rate [Baldonado, 2000; Hauptmann, 2001; Jeng, 2005; Orio, 2002]
- Friendliness [Bekele, 2002; Dillon, 1999]
- Learnability [Papadakis et al., 2002]
- Number of steps to take [Dillon, 1999]
- Personalization/customization (user's control; scalability / flexibility) [Bekele, 2002; Covey, 2002; Champeny et al., 2004; Papadakis et al., 2002]
- Precision/recall & variances [Browne, 2001; Jeng, 2005; Park, 2000]
- Responsiveness (adequacy of the system's response to users inquiries, [Dillon, 1999; Kengeri et al., 1999; Larsen, 2000])
- Time to complete/efficiency (actual and perceived) [Baldonado, 2000]
- Usefulness (for task in hand, meeting information needs) [Baldonado, 2000]
- Visibility of interaction status [Peng et al., 2004]

Having reviewed five U.S. based DL projects, Jones et al. (1999) concludes that "a digital collection can contain a critical mass of high quality, copyright-cleared content all organized around a solid metadata foundation, and still prove to be a failure." For this reason, usability is recognized as an indicator of DL quality and widely used for DL interface evaluation, whereby the primary objective is to examine the gap between system and user. Usability is defined by Chowdhury (1999, p.433) as "a system's capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and support, to fulfill a specified range of tasks, within the specified range of environmental scenarios."

By looking at the definition, one may see that usability can be measured from various aspects, such as *ease of understanding/use, error detection and handling, learnability, personalization* and *user control* (see Table 2.4 for a detailed list). In addition to these specific *usability measures,* conventional *effectiveness* (i.e.,

precision/recall or the variances) and *efficiency* (e.g., time to complete search task) criteria, as well as interface design principles (e.g., *simplicity, consistency, attractiveness,* etc.) are also employed in the evaluation of DL interfaces. Further, researchers and professionals have recognized the significance of taking into greater account the human-oriented factors in evaluating DL interfaces and the interactions between human users and interfaces. Frequently used user-oriented measures include *appropriateness to target users* in terms of fitting into their backgrounds, tasks and needs (Dillon, 1999; Zhang, 2004), *personalization and customization of interface functions and features* (Bekele, 2002; Covey, 2002; Champeny et al., 2004; Papadakis et al., 2002), *visibility of system status* (Peng et al., 2004), *assistance/help availability* (Huxley, 2002; Kengeri et al., 1999) and *friendliness* (Bekele, 2002; Dillon, 1999).

It is also worth noting that there are several DL interface studies that have been conducted in a more rigorous manner in terms of employing both subjective and objective measures (Hauptmann et al., 2001; Baldonado, 2000). To examine how efficient the proposed hi-cites result display mode (the combination of table and citation with use-controlled highlighting function) is when compared to solo table or citation display modes, Boldonado (2000) collected both *average time to complete tasks* and *perceived task-completion time*. Disappointingly, the paper does not report any comparative data on these two measures.

Likely, there exists more ready-to-use framework and/or criteria checklists for interface evaluation (e.g., Wesson's multiple view, Nielson's five measures and ten principles, Dillon's TIME framework, Mead & Gay.'s evaluation tool etc.) in contrast to the evaluations at the other five DL levels. Based on nine years of investigations of

human information usage from a human-computer interaction (HCI) viewpoint, Dillon proposes TIME (Task, Information model, Manipulation, Ergonomics) as a user-centered framework for DL interface evaluation. The framework suggests that a DL usability test should reflect users' contextual tasks, their cognitive representation of information space, and their information searching behaviors. In addition, an engineering consideration should also be applied. Whereas Dillon is essentially general without any suggestion of applicable metrics, Wesson (2002)'s views include more specific indicators for usability evaluation, including *diversity, complementarity, decomposition, parsimony, space/time resource optimization, self-evidence, consistency,* and *attention.* Compared to Wesson's multiple view in which the scientific basis for the list of criteria is unclear, Mead & Gay's interface evaluation tool is grounded on empirical data collection and analysis using a well-developed research method by Trochim in 1985 – Concept mapping (i.e., a structured conceptualization process relying on multivariate statistical analysis techniques). The tool is divided into four categories: *search/browse, search alternative, modification/sharing*, and *sorting/customizing*. Presumably, the sole inclusion of the function-focused categories lies in the exclusion of general users' opinions in the data analysis.

While these proposed DL interface evaluation tools have received little attention and application, Nielsen's usability test attributes and principles (1992; 1993) are well known and largely adopted in professional settings (Prown, 1999; Hennig et al., 2002; Peng et al., 2004). As a further extension of the five attributes (*learnability, efficiency, memorability, errors,* and *satisfaction*), the ten principles were formed on the basis of factor analysis on a set of usability problems and assumed originally to be more

appropriate for the involvement of specialists rather than general users (Nielsen, 1992). Nevertheless, by looking at several principles included, such as *match between system and real world* and *user control and freedom*, one may question whether or not specialists, the DL evaluation experts in Nielsen's usability test, are able to determine exactly the target users' "real world" and their information searching preference.

Still, the existing DL interface evaluations lack indicators specifically for examining unique DL features in terms of supporting social/group interaction among heterogeneous users (e.g., student, teacher, librarian, patron, physician, patient), utilizing hypermedia information, etc.

2.2.4 User Level Evaluation Criteria/Measures

Evaluations at this level primarily aim to examine individual user or user groups' outcomes after their DL use. The examination focuses on (1) changes of users' information behaviors, cognitive/decision-making/problem-solving capabilities, as well as affective differences; and (2) impacts/benefits as to users' task in hand, or later on research, work, life, etc. By looking at Table 2.5, one may find that there are essentially two types of criteria: output and outcome. Whereas output criteria (e.g., *session time, accuracy of task completion*) measure directly how well a given DL is used by a user or a community of users, outcome criteria are used to assess DLs in an indirect manner, either users' subjective opinions (e.g., *acceptance, intent to use, satisfaction*) or their post-use performance (e.g., *learning effects, productivity of users*).

The data on use/usage of a DL or a DL aspect are frequently collected via server logs. The measures vary among studies, including *number of login/sessions* (Brophy et al., 2000), *number of request per day/user/session/institution* (Marchionini, 2000),

instance of use (Abbas et al., 2002; Jones et al., 2000), *percentage of queries/searches* (Brophy et al., 2000), *document viewed* (Abbas et al., 2002; Bekele, 2002; Brophy et al., 2000, Shim, 2000), *repeated use* (Baldwin, C., 1998), *degree of penetration* (Bishop, 1998; Shim, 2000), and *user range* (Monopoli et al., 2002). Whereas the values of these measures can be obtained in an easy and direct way, Bollen and Luce (2002) propose a more complicated and subtle measure in terms of revealing *user community determined preferred relationships among documents and journals* from server logs. The Journal Consultation Frequency (JCF) was calculated by the sum of m_{ij} matrix, where m_{ij} corresponds exactly to the frequency of a co-retrieval event of two documents or journals. Not only can this measure be utilized to examine document usage patterns, the authors state, but also such relationships can be used as quality indicators of DL collection and service. Moreover, the authors used Impact Factor Discrepancy Ratio (IFDR), the ratio of JCF over IF (ISI's journal Impact Factor) to demonstrate the deviation of a specific user community from the general community in terms of journal impacts. Although the concept of JCF and IFDR is promising, it is doubtful how strong the two measures are as indicators for the user community's preference for collections and documents according to the following reasons. Firstly, there tends to be no empirical research that has validated the idea. Secondly, JCF tends to be more appropriate to serve as an association indicator among documents rather than a significance measure of these documents considering the "navigational lost" among hypertexts frequently reported by other researchers.

In addition to the objective criteria, a few subjective metrics can be found in the studies with respect to examining DL use, such as *acceptance* (Bollen & Luce, 2002; Mead & Gay, 1995), *intent to use/reuse* (Sumner & Melissa, 2001), *preference* (Bollen &

Luce, 2002; Park, 2000; Wildemuth et al., 2003), etc. Among these criteria, Sumner and Melissa's *reuse intent* is worth noting. After an interview with seven earth science faculty and two students who used DLESE (Digital Library for Earth System Education) in their teaching and learning, the authors found that *reuse intent* is one of five core factors associated with information use of these faculty and students. To a certain degree, continuous *reuse intent* may represent *user loyalty* as an outcome of using high quality library service, which was identified by Cullen (2001) from the review on the existing research on library ServQual—an ARL (Association for Research Libraries) project for libraries in the digital age.

Table 2.5: Existing Criteria for DL Evaluation – User

- Acceptance [Bollen & Luce, 2002; Mead & Gay, 1995]
- Accuracy of task completion [Wildemuth et al., 2003]
- Efficiency in terms of session time and redundancy of search procedure [Jones et al., 2002; Larsen, 2000; Meyyappan et al., 2004; Shim, 2000]
- Intent to use & reuse a collection or system feature [Sumner & Melissa, 2001]
- Learning effects [Borgman & Gilliland-Swetland, 2000; Greenberg, et al., 2002; Thebridge et al., 2002])
- Performance (of students) [Budhu & Coleman, 2002; Champeny et al., 2004; Borgman & Gilliland-Swetland, 2000]
- Preference [Bollen & Luce, 2002; Park, 2000; Wildemuth et al., 2003; Jones, 2000; Meyyappan et al., 2000]
- *Productivity of users* [Lyman, 1997]
- Satisfaction [Bishop et al., 2000; Bollen & Luce, 2002; Cullen, 2001; Wilson & Landoni, 2001]
- Use/usage [Abbas et al., 2002; Borghuis et al., 1996; Baldwin, C., 1998; Bekele, 2002; Bishop, 1998; Bollen & Luce, 2002; Brophy et al., 2000; Entlich et al., 1996; Jones et al., 2000; Khalil & Jayatilleke, 2000; Larsen, 2000; Marchionini, 2000; Monopoli et al., 2002; Shim, 2000]

Further, Cullen found that there might be a causal relationship between user satisfaction and loyalty. According to Saracevic (2000), *satisfaction* might contain material (e.g., desired information being found), cognitive (e.g., increased knowledge) and/or emotional (e.g., pleasant search experience) outcomes. In addition to subjective measurement of satisfaction through users' self reporting, the cognitive benefits of using a DL are assessed in an educational context through examining changes in students'

interest in learning, tests of information literacy skills in terms of analytic ability to assess the validity and reliability of information sources (Thebridge et al., 2002), and scientific thinking and reasoning skills (Borgman & Gilliland-Swetland, 2000). The improved learning skills are promising and essential factors responsible for final learning outcomes, such as concepts learned (Budhu & Coleman, 2002), graph understanding (Champeny et al., 2004), increased post-implementation (Borgman & Gilliland-Swetland, 2000), etc.. In addition to these context specific criteria, conventional effectiveness (e.g., *accuracy of task completion*) and efficiency (e.g., *time to complete tasks*) are still employed to assess the output of DL use.

In summary, most evaluations at this level focus on the use/usage and benefits in individual searching and learning. There is still room for identifying more appropriate criteria for assessing what changes are brought on by DL applications on users' daily work and lives. Evaluations at this level are essential for DL innovations, although they do not directly measure DL systems and their components. The rationale lies in Bishop et al.'s series of studies on DeLiver, a Digital Library Initiative (DLI) project at the University of Illinois, where their results demonstrate that insignificant barriers (e.g., trivial technical problems) "became magnified in the effect of use." Moreover, several studies have proved that there are gaps between DL developers' prediction of users' use and actual use by users (Blandford & Buchanan, 2002; Champeny et al., 2004).

2.2.5 Service Level Evaluation Criteria/Measures

The DL innovation does not mean that the roles of librarians diminish. Instead, library professionals have been facing growing demands and increased challenges in the digital age. Several studies show that DL users encountered difficulties and thus need

assistance in using and finding information (Blandford & Buchanan, 2002). Unfortunately, the role of librarians tends to be underestimated in current DL development and research domains (O'Day & Nardi, 2003).

Table 2.6: Existing Criteria for DL Evaluation – Service

- Accessibility–access barrier [Lankes et al., 2003; Cullen, 2001]
- Accuracy [Lankes et al., 2003]
- Cost [Brophy et al., 2000; Lankes et al., 2003]
- Cost-benefit [Kantor & Saracevic, 1999]
- Control capability [White, 2001]
- Courtesy [Lankes et al., 2003]
- Difference before and after service intervention [Bertot, 2003]
- Empathy [Cullen, 2001]
- Functionality [White, 2001]
- Gaps between expectations of users and providers [Bertot, 2003]
- *Personnel support* (e.g. number of staff engaged) [Kwak et al., 2002]
- Positive feedback [Carter & Jones, 2000];
- Reliability [Cullen, 2001]
- Responsiveness [Cullen, 2001; Lankes et al., 2003; White, 2001]
- Satisfaction (material/emotional) [Cullen, 2001]
- Use [Brophy et al., 2000; Carter & Janes, 2000; Cullen, 2001; Hauptmann et al., 2001; Kwak et al., 2002; Lankes et al., 2003; Shim, 2000]

For this reason, service evaluation is purposely separated from other evaluation levels for identification and discussion. Evaluation at the service level measures how well a DL can provide additional on-demand (in particular, human or human-like) assistance to users, such as reference, tutorials, term suggestion, active push service (e.g., SDD-selective document dissemination), etc.. Lankes et al. (2003) identified six criteria for evaluating digital reference (DR) from two studies in progress, namely *courtesy, accuracy, satisfaction, repeat users, awareness,* and *cost*.

In addition to the six metrics proposed by Lankes et al., there are several others worth mentioning (see Table 2.6). First, while *service awareness* can be used to evaluate digital reference accessibility, *initial expectation of convenience* (Lankes et al., 2003) and *actual difficulty in accessing the system* (Cullen, 2001) can be additional indicators.

Second, being different from face-to-face reference transactions, digital reference usually features time lag and invisibility in communication. As such, *responsiveness* (Cullen, 2001; Lankes et al., 2003; White, 2001) and *user's control* (White, 2001) become crucial. Also, conventional service quality indicators, such as *empathy* and *reliability (*Cullen, 2001*)*, should be applicable in the digital circumstance as well. Third, compared to single indicators, comparative criteria, such as *difference before and after service intervention* (Bertot, 2003), *gaps between expectations and users' perception* (Bertot, 2003), *the ratio of the number of repeat users over the number of potential users* (Borghuis et al., 1996), *cost-benefit* (Kantor & Saracevic, 1999), tend to be more convincing in demonstrating DL performance. Similarly, Kwak et al.'s (2002) *number of staff engaged in service* would be more meaningful given compared with total and/or potential number of users served. Fourth, whereas it is not natural and convenient to collect users' ratings on satisfaction, *the mount of positive feedback* (Carter & Jones, 2000) might have a similar meaning but in an unobtrusive manner. Fifth, *use/usage* is still employed widely with various measures for assessing digital reference outcome, including *number of active/repeat users* (Cullen, 2001; Kwak et al., 2002), *number of questions asked and/or answered* (Carter & Jones, 2000; Lankes et al., 2003), *number of user tutorial completed (*Kwak et al., 2002), *number of sessions per user*, *percentage of incomplete sessions over total attempted ones, percentage of the population reached by a digital reference service* (Brophy et al., 2000), etc..

2.2.6 Context Level Evaluation Criteria/Measures

A digital library usually has social and environmental dependency. On one hand, a successful digital library should comply with institutional/social practices; on the other

hand, it should be well supported by the institution and society within which it exists. Having acknowledged the significance of intellectual and financial roles in DL innovation, Blixrud (2002) and Lynch (2003) proposed using *sustainability* to measure the extent to which the augmentation of a DL could be secured without eventually losing its vitality. Similarly, in Besek's perspective (2003), DL developers should well observe intellectual property protection laws *(copyright protection).*

These proposed criteria or the like address how well a given DL fits into larger contextual (e.g., institutional, social, cultural, economic, legal) practices. Meanwhile, some other criteria have been suggested or used to examine what impact and effect DLs may have on these contextual practices. As pinpointed by several leading researchers (Bishop, 1999; Marchionini, 2003; Saracevic, 2000) in the domain of information science, DL effects are not only restricted in the ways in which people find information, but also the ways in which they live in society. Specifically, as seen from existing practices, DL innovations have yielded to tremendous changes in a variety of fields, such as publication, education, research, healthcare, business, entertaining and daily living. Bishop (1999) compares the information use of two different groups of users (i.e., academic and low-income communities), and shows that DL usage is associated with social practice, beliefs/goals, community norms, knowledge, technology access/proficiency, resource constraints, and the interplay among them. Unfortunately, there is no evaluation in the study regarding how DLs change the daily lives of these two communities.

To date, only a very few evaluations have to some extent examined contextual effects of DLs (see Table 2.7), among which evaluations in educational settings take up more counts. Two studies (i.e., the Perseus Digital Library by Marchionini et al in 2001

and the Alexandria Digital Earth Prototype by Borgman & Gilliland-Swetland in 2000) can serve as two exemplars demonstrating how DL evaluation can be done at this level. Whereas Borgman et al.'s evaluation primarily employed both classroom observation and focus group for classroom learning style, Marchionini et al. combined a greater variety of research methods, including observation, focus group, as well as document analysis and learning analysis reported by instructor, for the single research purpose of examining educational impacts of Perseus. Sadly, neither evaluation has produced convincing indicators, which can demonstrate how teaching/learning styles have been shaped/reshaped by the use of DLs in classrooms.

Table 2.7: Existing Criteria for DL Evaluation – Context

- *Outcome against predefined goals* [Bertot & McClure, 2003; Star, S.L. et al., 2003]
- Copyright compliance* [Besek, 2003; Jones et al., 1999]
- Local vs. remote use [Shim, 2000]
- *Organizational usability* [Elliott, M. 1995; Kling and Elliott, 1994; Xie & Wolfram, 2002]
- Productivity of community members [Lyman, 1997]
- *Sustainability* (of current and augmented collections) [Blixrud, 2002; Lynch, 2003]

* "copyright compliance" was termed as "copyright abidance" in the instruments, including the interview cards, the survey forms, and the post-search questionnaire in the experiment.

The scarcity of contextual evaluation might be associated with measurement hardship at this level. Nevertheless, several researchers' depictions and arguments are promising and heuristic. While advocating the potential impact of information technology on scholarly communication, Lyman (1997) pinpoints that social functions "are not easily measured in terms of outcomes, but are an element in the *productivity of faculty and students*." Bertot & McClure (2003) propose that library outcomes in the digital age can be measured by the *extent to which a given library /service meets the predefined goals by the library and/or anticipated by the community the library serves* (e.g., academic

institution, county, city). The proposed measurement shares the meaning with Star et al. (2003)'s notion of *convergence between information artifacts and communities of practice* for the purpose of "transparency beyond the individual level of scale" in a socialized digital world and Kling and Elliott (1994)'s "design for *organizational usability*." As opposed to interface usability, where features and contents on a given interface are examined in terms of the extent to which they can assist individual users to find needed information, the *organizational usability* is devised to examine how well a DL system is "integrated into the work practices of organizations." Furthermore, Kling and Elliott suggest employing anthropology research methods to identify the work practice based on a literature review and empirical observation.

In general, the literature reviews reveals that DL evaluation at this level deserves and demands much more studies with respect to (1) identifying more appropriate criteria in different user communities perhaps through beginning with examining each community's expectations and goals towards DL innovation; (2) extending evaluation objects beyond education, research and business to people's daily lives; and (3) conducting more actual evaluations at this level.

2.3 Methodologies

2.3.1 Benchmarking Evaluation Methods and Criteria

In addition to evaluation framework, criteria and measures, an appropriate methodology can also influence evaluation outcomes. The criteria listed in the Table 2.2 through Table 2.7 imply that both quantitative and qualitative research methods are essential for DL evaluations. *Qualitative research methods* are most frequently used in evaluation studies at interface, user and environmental levels. The methods include

interviews (Borgman, 2000 & 2001; Hauptmann, 2001; Marchionini, 2001; Sumner, 2001), *focus groups* (Hill, 1007 & 2000; Marchionini, 2001), *ethnographic observations* (Borgman, 2000 & 2001; Hill, 1997 & 20000; Khoo, 2001; Marchionini, 2001; Seadle & Peters, 2000), and *content analysis* (Marchionini, 2001). By contrast, *quantitative methods* are much employed to evaluate system performance and the outcome of digital information organization and representation. The most frequently used methods are *log analysis* (Abbas, 2002; Carter, 2001; Jones, 2000; Sfakakis & Kapidakis, 2002) and *experiments* (Baldonado, 2000; Hauptmann, 2001; Hee; 1999; Hidaka, 2001; Ma, 2001; Park, 2000; Purcell et al., 1997; Rui, 2000; Sumner, 2001, Zhang, 1995). *Survey* (Hill, 1997 & 2000), as a special research method somewhere between the qualitative and quantitative approaches, is also widely used.

A number of studies have combined various qualitative and quantitative methods (Bishop, 2000; Borgman, 2000; Dorward et al., 2002; Entlich et al., 1996; Hill, 1997 & 2000; Van House, 1996; Kassim & Kochtanek, 2003; Marchionini, 2000; Spink, 1998). It should be noted that many quantitative lab studies (Baldonado, 2000; Park, 2000; Sumner, 2001) employ special research techniques, such as *thinking aloud protocol*, to collect subjective data from users. The combination of different research methods is promising in terms of being able to increase the validity of evaluation findings while providing rich data for assessing DL outcomes. Mead and Gay (1995) innovatively adapted Trochim (1985)'s concept mapping in their evaluation of MoA (The Making of America) at Cornell University. Whereas brainstorming elicited the epistemological orientations of different stakeholders, the statistical concept mapping was used to compare and synthesize various sources of data.

Wilson et al. (2001) outline a framework for evaluating electronic books to standardize the research of this kind and ultimately increase the comparability of findings. The framework includes four main methodological aspects for e-book evaluation, namely selection of material (e.g., collection), selection of actors (e.g., participants, evaluators, task developers, and task assessors), selection of tasks, and selection of evaluation techniques (e.g., questionnaire, behavior observation, think-aloud, and interviews). Although these four aspects are critical to evaluation results, there should be more (e.g., selection of criteria) that are of equal importance. Additionally, the framework seems to be useful as well to other DL settings, not only to electronic books.

In addition to methodological standardization, some researchers have been working on developing generic evaluation schemes/models for benchmarking. Among these studies, only a few provide criteria for multiple dimensions of DL evaluation, such as Kwak et al.'s evaluation model in 2002 and US Digital Library Initiatives Metrics Working Group's quantitative performance measures by Larsen in 2000. The others are primarily proposed for a single level of evaluation. For instance, whereas Dillon's TIME framework (1999), Mead & Gay's evaluation tool and Wesson's usability evaluation indicators are proposed specifically for interface evaluation, Saracevic & Kantor's taxonomy of library and information service value is constructed for library service value assessment. Each provides more or less specified criteria for DL evaluation except Dillon's TIME (Task, Information model, Manipulation, Ergonomics) framework (1999), in which only four dimensions of interface evaluation are proposed.

2.3.2 Developing Evaluation Model for Benchmarking

Not only should attention be given to what evaluation framework has been proposed, one also needs to know how they have been developed in order to see whether a given framework is valid and generalizable. Among the handful of studies on evaluation framework construction, two schools of research approach are identified. One is top-down, the other is bottom-up. The latter starts by examining perspectives/opinions of DL stakeholders through interviews, focus groups, brainstorming, and surveys, and then generalizing them using statistical techniques. Kwak et al. (2002) developed an evaluation model for a university library via two-phase research. In the first phase, an initial model was constructed based on the opinions of library experts and the previous works on the evaluation of both traditional and DLs. In the second stage, Thrice-run Delphi surveys of 50 DL-related professors, researchers, and university librarians were applied to develop a valid evaluation model. Eventually, a new model was finalized with 7 categories (goal setting/vision, library specialization, information resources, information usability environment, information sharing, information services, and human resources & budget), 35 items, and 92 indicators. Although the model is enlightening in terms of including a large portion of criteria for environmental effects, such as *library plan, library specialization, information sharing, information service*, and *human resource,* the validity of the model is doubtful considering the exclusion of the opinions of library users.

Comparatively, Mead and Gay (1995)'s evaluation framework for MoA (The Making of America) at Cornell University tends to be more valid, since the model construction takes into account perspectives via brainstorming from more diverse DL

stakeholders (i.e., development staff, graduate students, university staff and professional searchers). Similarly, when Saracevic & Kantor (1997) were developing their taxonomy of library and information service value, they conducted focus groups, surveys, and interviews with users of 18 services in 5 research libraries.

In contrast to the bottom-up approach, the top-down starts by utilizing existing DL constructs, evaluation criteria, or the researchers' own perspectives, and ends up with a test of the proposed framework. White (2001) proposes a descriptive model for analyzing and evaluating digital reference services with approximate 100 questions from18 categories of four broad areas. To test the model, White analyzed 11 digital reference services. The results show the usefulness of the model while illustrating strengths and weaknesses of each service in the aspects examined. Fuhr et al. (2001)'s evaluation scheme for the DELOS working group on "the DL test suite" was formed through an examination of the combined DL notion from both research and professional domains. Then, two sets of Web-based surveys primarily of DL developers and researchers, with one for existing DLs and another for future DLs, demonstrate that the proposed scheme "seems to be appropriate for DL characterisation". The scheme covers four DL aspects: *data/collection, system/ technology, users* and *usage,* and each aspect includes a list of major attributes. The framework is promising in terms of integrating viewpoints from both DL developers and researchers. However, its holistic value might be weakened for the following three reasons. First and again, the framework doesn't take into account the viewpoints from users—one of the key shareholders in DL innovations. Second, it is unwise to leave out interface, a crucial component in any information retrieval system where system and its users meet and communicate. Third, environmental

effects, which are highlighted by a number of researchers, are not included. Similarly Dillon's TIME (1999) is grounded on the authors' nine years of investigating of human information usage from a HCI viewpoint. Although the author claims that the framework is from user's perspectives, the validity of the framework tends to be weak, since there is no supportive empirical evidence.

With respect to evaluating libraries in the digital age, three other large-scale programs are worth mentioning: ARL's E-metrics, UK's eVALUEd, and EU's EQUINOX. All three aim to develop generic evaluation models for libraries in the digital age. The eVALUEd's evaluation tool kits (http://www.evalued.uce.ac.uk/) were initiated with a survey of higher education institutions in UK with particular focus on what evaluation techniques were being employed, who would use evaluation results, how the results might affect decision-making and what evaluation could be conducted given more time, resources, staffing etc. Similarly, EQUINOX (http://equinox.dcu.ie/) and New Measures Initiatives (http://www.arl.org/stats/newmeas/index.html) started with an inventory of current library performance measures. Additionally, these three projects all end up with more or less detailed checklists of indicators. The working group members for the three projects are all primarily professionals in librarianship. By looking at the research procedures as summarized in Table 2.8, one may find that none of the projects has incorporated users' opinions in the composition of its performance indicator checklist. Additionally, whereas ARL's New Measures Initiatives and UK's eVALUEd embrace outcome measures at the institutional level in terms of evaluating libraries in the digital age within the higher education environment, EU's EQUINOX metrics are limited in the

traditional library performance scope. Furthermore, to date, no reports have been found with evidence demonstrating how good these metrics are.

Table 2.8: Comparison of eVALUEd, E-Metrics, EQUINOX

	Research Procedure	Product Outcome
eVALUEd	Start with a survey on higher education institutions in UK, then move onto follow-up interviews with a selection of survey participants, interviews with experts in the field of evaluation generally and EIS specifically, and review of the literature and related projects. Major sources will be included in the online toolkit. Finally, end up with test-bed work.	A set of checklists of measures as well as corresponding instruments towards four dimensions: use/usage, access, management, and resources. There are two types of instruments: one for librarian and another for students
New Measures Initiatives	Three phrases: (1) inventory of current practices at ARL libraries as to statistics, measures, processes, and activities that pertain to networked resources & services, (2) identification and field testing of statistics & measures, recommendation of measures through surveys and onsite test (to date, 49 institutions participate in the test), and (3) identification of linkage to educational outcomes and impacts, to research, and to technical infrastructure through content analysis on standards from education commissions.	The indicators are grouped towards four dimensions for assessing e-resources (E-metrics), DL services (E-Qual), regular library service (LibQual), and learning outcomes. Among which, E-metrics contain 16 indictors including R_1 - R_3 (number of e-resources), $U_1 – U_5$ (use/usage), $C_1 – C_5$ (cost), $D_1 – D_3$ (digitization activity), E-Qual is for the NSDL project
EQUINOX	Three phases: (1) investigate existing library performance measures, (2) incorporate with technical perspectives of automatic system design, & (3) validate through testing by a large number of participating libraries (Brophy et al., 2000)	A list of (14) electronic library performance indicators with respect to use/usage in system and service, cost, staffing, and users' satisfaction

As a core DL research activity in the US, Digital Library Initiatives (DLI) also established a Metrics Working Group to develop quantitative performance measures for system, user and content aspects based upon synthesizing the results from six test-beds (Larsen, 2000). In the summative paper, Larsen reports the research findings: a DL can be measured with two IR phrases: query and retrieval. For each of these, four criteria should be examined: (1) *timeliness* (speed), (2) *sufficiency* (the adequacy of the system's response to queries), (3) *correctness* (precision), and (4) *effort* (the amount of work required by the user to interact with the system). Meanwhile, seven dimensions of DL

evaluation options were identified. These findings are not promising in terms of being able to evaluate a DL in a holistic manner, presumably because of the predetermined evaluation foci, the small size of test-beds, and exclusion of broader DL contexts other than content, system and user.

Geared specifically towards efficiency of research libraries in the digital age, Shim & Kantor (1999) propose a Data Envelopment Analysis (DEA) analysis tool in which inputs are collection, staff and university characteristics, such as total volumes, number of staff and FTE students, and outputs are activity measures such as number of ILL transactions, items circulated and reference transactions. In line with Lancaster's cost-effectiveness measure, The DEA approach is enlightening and heuristic in terms of being able to evaluate a DL in a more convincing manner.

In sum, as described by a number of researchers (Chowdhury & Chowdhury, 2003; Marchionini et al., 2003; Saracevic, 2000), DLs can be evaluated at various levels and elements by using different criteria and research methods. However, the literature review reveals that, in general, DL evaluation is still at an immature stage. The depiction is based on the following facts: (1) the evaluation efforts have not caught up to the pace of DL development. Most DL projects lack evaluation reports; (2) a systematic evaluation framework is not widely adopted for each DL project. Only a very few DLs have been assessed at different levels with various viewpoints; (3) there is no promising generic framework for DL evaluation. None of the proposed frameworks can serve as a holistic model for DL evaluation, given their limited DL aspect coverage, inadequate diverse stakeholders' involvement; and (4) existing criteria are inadequate to reflect the complexity of DLs. The majority of criteria found is primarily invented and/or favored by

professionals and researchers in the domains of librarianship, information retrieval, and human-computer interaction. To improve this immature situation, one promising action would be to develop a generic and holistic DL evaluation framework that includes a set of criteria. These criteria should capture core DL aspects and characteristics while embracing perspectives from a wide variety of stakeholders.

Chapter 3 RESEARCH OBJECTIVES AND THEORETICAL FRAMEWORK

3.1 Research Objectives

The main purpose of this research is to develop such a holistic DL evaluation model. The expected framework will have the following three meanings in terms of being holistic: (1) to cover all DL levels, including digital information, collection, software, hardware, interface, service, user, as well as context; (2) to bring in perspectives from as many diverse groups of stakeholders as possible; and (3) to be open to different kinds of evaluative indicators, no matter whether they measure input, output, or outcome, and no matter whether data features are qualitative or quantitative. To develop such a framework, the research objectives are:

1. To identify what criteria can and should be used in DL evaluation and construct a preliminary set of criteria for different DL levels through examining existing studies and eliciting DL experts' opinions;
2. To examine at a large scale how important each criterion in the preliminary set is in the perspectives of more diverse stakeholder groups, and build a model in which criteria perceived to be "important" are presented in a meaningful manner;
3. To test the validity of the model when it is applied to real-life DL evaluation.

Whereas the first objective is examined primarily through a literature review and a semi-structured interview, the second via an extensive online survey; the third through evaluations on a university library Web site with the participation of different groups of stakeholders.

The three studies are carried out in three consecutive stages. In spite of focusing on different objectives, they are conceptually and methodologically interrelated. The

design of the follow-up studies relies on the previous research findings. For instance, the criteria identified from the literature review are used as a set of probes in the interview. Similarly, the survey questions and the criteria used in searching the library Web site depend on the research findings from the exploratory and confirmatory stages respectively. In turn, the findings from the follow-up studies may serve to verify the previous findings. Moreover, the construction of the holistic evaluation model is grounded on a representative spectrum of research methods with qualitative methods (literature review and interview) at the start and quantitative approaches (survey and experiment) at the end. The selection and sequence arrangement of these research methods are carefully planned with consideration of being appropriate to corresponding research objectives as well as maximizing the strengths of each method. Whereas the combination of literature review and interview tends to identify a more comprehensive set of criteria, the use of the quantitative methods with the involvement of larger numbers of and more diverse stakeholders and the support from statistical inferences has the strength to increase the validity of the final model.

The holistic framework based on pursuing these research objectives is able to serve as one of the most comprehensive models for DL evaluation, and ultimately contribute to the DL research and development, for the following reasons: (1) being in light of two promising conceptual DL evaluation frameworks (see the succeeding section for details), that is Marchionini (2000; 2003)'s multifaceted approach for inclusion of various groups of stakeholders in the research and Saracevic's (2000) stratified standpoint for categorizing DL levels; (2) relying on complementary research methods (i.e., semi-structured in-depth interview, extensive online survey, and experimental

evaluation) for the three research stages of which the research purposes and instruments are interrelated; (3) being grounded in the perspectives from various stakeholder groups (i.e., administrators, developers, librarians, researchers, and general users); and (4) considering both qualitative and quantitative indicators. This framework will contribute to the improvement of the chaotic status of DL development and research with respect to providing DL evaluators a comprehensive set of criteria to tailor for many purposes, as well as informing the domain of DL research and development in a more rigorous way what general users expect from a DL, which might be different from theirs as being addressed in several studies (Blandford & Buchanan, 2002; Cullen, 2001).

Having identified from the previous literature review the two weaknesses in the existing DL evaluations (i.e., criteria for examining contextual effects and from users' perspectives), this research strives to overcome the weaknesses. Specifically, the online survey and the experiment collect data from various groups of people, including researchers, professionals, as well as general users. Meanwhile, considering the scarcity of context-focused evaluation criteria, several interview questions are devised as probes to elicit thoughts and opinions on how a DL may influence institutional, social, cultural, legal, economic, and/or educational environments (see Chapter 4 and Appendix 3 for details).

3.2. Theoretical Frameworks

There are two theoretical frameworks in light of which the research is designed and conducted. One is Marchionini's multifaceted approach (2000; 2003) for assessing DL impacts. Another is Saracevic's stratified model (1996; 1997; 2000).

3.2.1 The Stratified IR Model and DL Evaluations

The stratified information retrieval (IR) model was proposed by Saracevic in 1996 & 1997. The model views an IR system as an entity containing components at different levels: informational, computational, interface, query, user, situational, and environmental. The IR system functions through interactions among these stratified components. The stratified IR model addresses essential components of an IR system in a more systematic and comprehensive manner. The interactive IR model is well suited for a holistic evaluation of IR systems. Meanwhile, each level in the model can be used separately to assess a system at a particular level. The comprehensiveness and flexibility can be regarded as two strengths of the interactive IR model.

The stratified model opens a new avenue for DL evaluations. In light of the model, a DL can be assessed at a single level or at two or more levels. Based on the model, Saracevic (2000) proposes a conceptual framework for DL evaluation that outlines the constructs and contexts of DL development. The components listed at each level are used as the constructs of a DL, whereas each stratum serves as the context within which an evaluation can be conducted. Once both contexts and constructs are determined, the DL evaluation is likely to have a clear and definite direction.

3.2.2 The Multifaceted DL Evaluation Approach

Whereas Saracevic's stratified model suggests what can be evaluated, it tends to be weak in providing a clear guideline regarding how quality data at stratified levels can be collected, analyzed and reported. Marchionini's multifaceted DL approach is likely to be a complementary framework. Having nicely tackled the complexity of DL development in terms of involving various people and activities, the multifaceted approach pinpoints that DL evaluations should be conducted through collecting data by taking different viewpoints, using different approaches and from different dimensions, then integrating data, and finally reaching a conclusion. Their integration must be "systematic, interpretive, and driven by high-level goals." Additionally, to make a DL evaluation multifaceted, it is essential to examine information needs from different groups of stakeholders, including users, librarians, administrators, developers, researchers, funders, etc., as well as tasks arising from their needs.

In light of the multifaceted approach, a holistic evaluation framework can be developed by incorporating perspectives from diverse groups of stakeholders with foci on their perceptions of the existing DLs, their expectations from future DLs, and how they would judge the quality of a DL.

The stratified and multifaceted approaches are enlightening in terms of having provided general guidelines for developing a holistic DL evaluation framework in which various people's perspectives towards all kinds of DL levels are included. In turn, such a holistic framework can specify the ideas embedded in the two general approaches. Furthermore, the process of and the data underlying the development of such a holistic

framework might be used as empirical verification to the two general DL evaluation approaches.

Chapter 4 METHODOLOGY

To develop the holistic DL evaluation model, a hybrid research approach combining both qualitative and quantitative methods is applied. Specifically, a three-stage research approach--*exploration, confirmation* and *verification (See Figure 4.1)* -- is devised to identify as many and as diverse as possible the criteria that can and should be used in DL evaluation, and eventually to construct a valid model with the inclusion of promising evaluation criteria. The promising evaluation criteria should be those perceived as being important* in various stakeholders' perspectives. During the *exploration stage*, a semi-structured interview and a representative literature review are employed to collect perceptions from DL experts (administrators, developers, and researchers) with a particular emphasis on determining what criteria can and should be used in DL evaluation. Then, the criteria identified from the interview and the literature review are embedded into an online questionnaire during the *confirmation stage*. Various groups of DL stakeholders, including administrators, developers, librarians, researchers, and general users, are asked to rate the importance of each criterion on a 7-point Likert scale. Additionally, the questionnaire is open-ended to collect additional important criteria from the survey participants. The holistic model is constructed by using some descriptive and inference statistical techniques.

* To differentiate the latter from the statistically "significant" from later on ANOVA testing results, "important" (or "importance") was used in the text body to replace "significant" (or "significance") in the original research instruments.

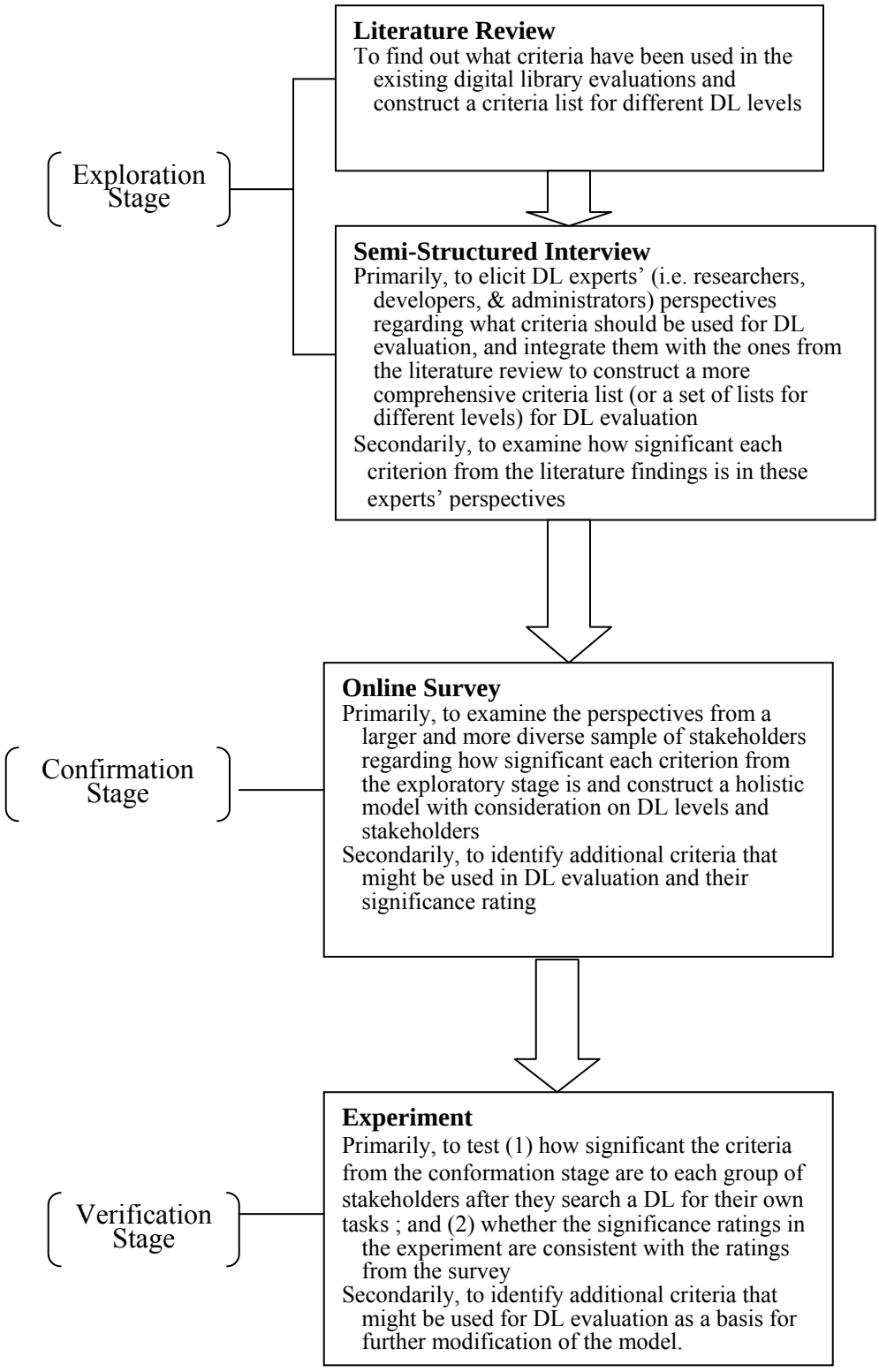

Figure 4.1: Illustration of the Three-Stage Research Approach

The selection of research methods is made with a consideration of being appropriate to corresponding research objectives as well as maximizing the strengths of each method. As clearly stated by Auerbach & Silverstein (2003), a qualitative research method is useful when researchers' knowledge about a targeted problem is limited, and can be used as "a tool for studying difference between groups" (p.26). The interview technique is good for eliciting a person's tacit thoughts, in particular when he/she has rich knowledge of the topic (Auerbach & Silverstein, 2003; Lindlof, 1995). Therefore, it should be appropriate to start with a semi-structured interview in developing the holistic DL evaluation model in order to address the problems arising from the literature review findings regarding (1) the lack of DL specific evaluation criteria and (2) uncertainty about the divergences among different DL stakeholder groups. Considering that personal knowledge varies among individuals, especially among DL stakeholders with different backgrounds, it is necessary to involve as diverse stakeholders as possible as interview participants. As suggested by Auerbach & Silverstein (2003), interview participants in qualitative research should have a rich knowledge of the research topic; accordingly, DL researchers, developers and administrators are selected as the interviewees in the study because the author believes that these three have rich knowledge and many experiences with DL related topics, and thus should be able to contribute DL evaluation criteria. Meanwhile, the literature review findings may serve as the basis and probing points in the interview for eliciting more in-depth knowledge and experiment of the three groups of DL stakeholders.

Whereas the interview may yield a rich set of DL evaluation criteria from a few DL experts' perspectives, the online survey is more appropriate for confirming the

importance of these criteria through the perspectives from more respondents and more diverse groups of DL stakeholders. Additionally, as a research method with both qualitative and quantitative features, open-ended questions in the survey form can be used to enrich the criteria set, while statistical inference increases the validity of the final evaluation model. Furthermore, an online survey is economic and effective in terms of reaching more survey participants at a lower cost.

No model can be considered perfect without being tested in a real application setting, including the anticipated holistic evaluation model. Accordingly, evaluation testing with stakeholders' interaction with a real digital library (the Rutgers University Library Web site in this study) was conducted to examine the model when it is applied to a genuine setting. The author anticipates that not only can the findings from the evaluation testing verify the validity of the final model, but they might help refine the model.

Although there is a debate regarding whether a library Web site can be considered a DL, by comparing typical features on a representative library Web site (e.g., Rutgers University Library) with the DL definition proposed by the Digital Library Federation (DLF), one may see that the former is essentially comparable with the latter.

> Digital Libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.
>
> — DL Definition by the DLF (Waters, 1998)

Specifically, the Rutgers University Library (RUL) site at http://www.libraries.rutgers.edu can be seen as a Web presence of the library because the site contains a clear statement about the *organizational* mission, a well defined *user*

community, as well as a presentation of its *organizational structures* and resources. Specifically, the site provides RU students, faculty and other RU *affiliated community members* with *readily and economically available resources* that are *selected*, *organized and integrated*, as well as *maintained* by *specialized staff*. From the site, faculty and students not only can search for physical library *collections*, but also *access digital works (*e.g., online full text*).* Meanwhile, they may readily seek online *intellectual* assistance from *specialized librarians*. From the author's standpoint, a qualified DL does not necessarily require all digital works reside on a single local server. Instead, DLs should fully take advantage of network technologies, that is to integrate distributed resources from different digital depositories.

4.1 Literature Review – The Exploration Stage

The literature review was performed in 2003 - 2004 with the following four primary approaches: (1) identifying and selecting DL related sources that are likely to cover DL evaluation literature; (2) constructing search statements and composing search queries to retrieve more DL evaluation literature; (3) selecting papers from retrieved sets that have representive contents about DL evaluation frameworks, methodologies, criteria and measurements; and (4) summarizing the frameworks, methodologies and criteria from the papers selected.

4.1.1 Identification of Sources

Various DL related sources (i.e., databases, journals, proceedings, Web sites) were searched to discover criteria that have been used or proposed in existing research and development. Whereas several key databases in the field of LIS (i.e., *Library & Information Science Abstracts, Information Science Abstracts, Library Literature &*

Information Science, ACM Digital Libraries, and *IEEEXplore)* were the starting points of the search, the majority of papers that have been selected are from some leading LIS journals (e.g., *Journal of American Society for Information Science and Technology, Information Processing and Management, Journal of Documentation*) and proceedings (e.g., *ACM/IEEE-CS Joint Conference on Digital Libraries, European Conference on Research and Advanced Technology for Digital Libraries*). Additionally, DL specific journals (e.g., *D-Lib Magazine, Journal of Digital Information*) as well as DL project Web sites (e.g., *Digital Library Initiatives, ARL E-Metrics, EU EQUINOX, UK eVALUEd*) also served as core sources to identify DL evaluation criteria.

Considering the breadth of DL influences, Web of Science (WoS) was examined to expand the search scope from the dominant body of DL research and development to the plausible DL application domains (e.g., education, health, etc.). WoS is a multidisciplinary database that indexes research from leading journals in each discipline. Therefore, the WoS search results should be representative rather than comprehensive.

There are mainly two reasons for combining the two literature review sources. First, to examine the state-of-the-art in DL evaluation, it is critical to review the work by people in the LIS field, including information scientists, library professionals, and computer specialists, who are more likely to conduct and report evaluation research. Essentially, the literature review needs to be done in a way that reflects the multidisciplinary and collaborative nature among these scholars and professionals. Second, considering the multiplicity of DL application domains and decentralized distribution of disciplinary resources, it is necessary to extend the search to other domains, such as education, business, etc., although it is unrealistic to perform a comprehensive

search in all DL applicable areas. Also it might be difficult to locate DL evaluation research in the areas, because people from other disciplines might have alternative terms for DLs, such as “online resource”, “Web-based resource”, “interactive teaching material”. It is impractical and unnecessary to create an exhaustive set of cross-disciplinary literature.

In short, the selection of representative literature focused on dealing with DL evaluation framework, criteria, and/or methodology. Eventually, about 155 papers were identified and selected as the literature review resources.

4.1.2 Search Query Composition

The primary search statement for the database search was formed by a Boolean logic combination of *(digital library or electronic library) and (evaluation or assessment or performance or outcome).* However, specific search queries varied among databases in accordance with a given database requirement with respect to truncation, phrase presentation, Boolean logic combination, etc.

4.1.3 Paper Selection

The papers selected for the review were restricted to those studies with representative thoughts and/or achievements on the frameworks, methods as well as criteria for DL evaluations. Eventually, about 155 papers were selected for having met the requirement.

4.1.4 Criteria Summarization

The literature review placed more effort on identifying criteria in the selected papers, focusing on the objective of the dissertation research, that is, what criteria can and should be used for DL evaluation. The six tables displayed in the literature review

chapter (Table 2.2 to Table 2.7) are the results of this procedure. Each table contains representative criteria for a given DL level, ranging from content (i.e., object, information, meta-information, and collection), technology (i.e., hardware and software), interface, service, user to context. The criteria lists served as the basis for developing an interview protocol in the next research stage.

4.2 Semi-structured Interview – The Exploration Stage

4.2.1 Interview Participants

A purposive sampling method was employed to select nine DL experts, who were more likely to provide insightful thoughts and opinions about DL quality/performance indicators. Three groups of interviewees (i.e., administrators, developers, and researchers) with three in each group participated in the research. These interviewees were recruited from the School of Communication, Information, and Library Studies and the university libraries at Rutgers University. A brief solicitation letter (see Appendix 1) was sent to candidate interview participants via email. Interview participant eligibility required substantive knowledge of DLs and/or adequate experience developing, administering or conducting research on DLs. Specifically, an eligible DL researcher should be one who has published at least one paper on DLs, and/or at least taught one DL course. Meanwhile, an eligible DL developer should have participated in at least one DL project by either designing or implementing the DL. In contrast, an eligible DL administrator does not have to be involved in detailed DL development, while his/her primary role is to overall control of the implementation of at least one DL or one aspect of a DL.

The rationale for inclusion of these three groups of DL community members lies in Saracevic's Stratified IR Model (1996; 1997; 2000) and Marchionini's multifaceted

DL evaluation perspective (2000). According to these two DL researchers, a DL contains a number of components at different levels, and its application involves the activities and interests of various stakeholders. Each DL level or activity is usually associated with a certain group(s) of people. For instance, the work of DL developers tends to focus on computing, engineering, and interface levels, whereas administrators are more likely to highlight situational and environmental effects. Compared with the ones from the professional DL community, scholars from the academic DL domain tend to have a more general vision regarding what has been done and what needs to be done, but they might have less pragmatic experiences. In other words, different DL stakeholders are likely to be able to provide perspectives on various components and levels. In order to develop a holistic evaluation framework for a DL, it is essential to collect and analyze the perspectives from as diverse groups of stakeholders as possible. Although general users are key DL stakeholders, they are unlikely to have acquired a deep knowledge/experience of DLs and therefore are excluded as candidate interview participants.

4.2.2 Data Collection

During June to October 2005, a semi-structured interview was conducted to collect the nine DL stakeholders' perspectives on DL criteria. Each interview participant was interviewed once by the author for about an hour. Appendix 3 shows the semi-structured interview protocol. The interviews were taken in places where the interviewees felt comfortable. At the beginning of each interview, the author informed the interviewee of the purpose and expected benefits of the study, and asked him/her to read and sign an interview consent form both for permission to be interviewed and to be audio-taped (see Appendix 2). Via the consent form, the interview participants were informed that there

would be no connection between the data collected from them and their identity. The interview protocol had been approved by IRB (Investigation Review Board) at Rutgers University.

The original interview protocol was pilot tested in early 2005 by interviewing three DL experts from Rutgers University. In general, the pilot was effective with respect to (1) having examined the validity of the exploration stage design; (2) having identified some constructive aspects for the research design in the succeeding stages; and (3) having facilitated forming an initial coding scheme. The interview protocol proved to be valid by examining the interviewees' responses to the questions and the pilot results. There were two small changes that needed to be made. First, a few criteria on the cards needed to be defined more clearly and more precisely. Second, the question for evaluation criteria of the DL as a whole was removed to ensure the interview to be completed within an hour. Additionally, the question was deemed unnecessary because it generated largely repetitive narratives captured by other questions.

The nine questions in the interview protocol (see Appendix 3) were asked in the same order so as to minimize instrument bias. These questions were grouped into two sections: experience and personal perspective. The experience related questions were used to collect interviewees' background with DL research, development administration, and/or evaluation, as well as their related gains, lessons and thoughts. The questions were comparatively easier to answer and could be used to refresh the interview participant's thoughts. Additionally, the background information from the experience recounted should be useful to generate more insightful viewpoints for the questions in the second section. Meanwhile, a question (i.e., what comes to your mind if you are asked to provide

a definition?) was used to elicit interviewees' thoughts about DL notion, considering a plausible association between individualized views of DL notion and personal perceptions on important evaluation criteria. The second section was utilized to collect the interview participants' perspectives regarding what criteria could and/or should be used in DL evaluation. Each question in the section was targeted a given DL aspect, ranging from content, technology, interface, service, user to context (see Appendix 3 for the detailed protocol).

For each DL aspect, in addition to the question-answer (QA) during which the interviewees were asked to speak freely about what criteria should be used for DL evaluation in their perspectives, a card sorting (CS) technique was employed to have them rank 8-10 criteria for each DL level evaluation. The criteria included on the cards were those pre-selected by the author from the literature review results (i.e., criteria in Table 2.2 through Table 2.7 that were used and/or proposed by others for the given level of DL evaluation). All criteria pre-selected were those with higher frequency of use and/or recommendation. All interviewees were asked to sort these cards based upon their perceived importance of each criterion to DL evaluation at the given level. During the card sorting, an interviewee was encouraged to refer to the back of a card for the definition of the criterion (see Figure 4.2 and Figure 4.3 for a sample sorting card). The Appendix 3 contains a detailed list for individual criteria and notions.

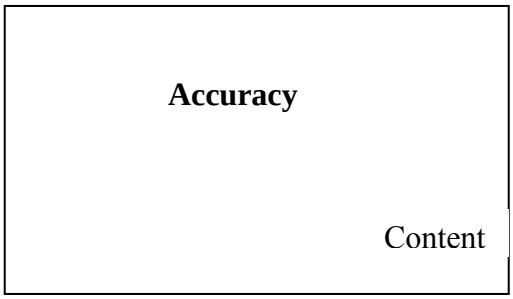

Figure 4.2: Sample Interview Card (Recto)

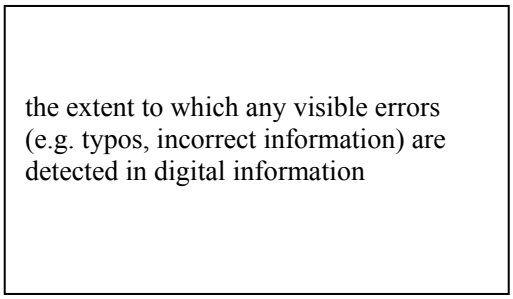

Figure 4.3: Sample Interview Card (Verso)

When an interviewee was verbalizing his/her thoughts, the author occasionally used appropriate probes to encourage him/her to articulate more thoughts and opinions. These techniques included: (1) asking neutral questions, such as "anything else?", "could you please tell me more about it?", "what do you mean by…?"; (2) repeating an interview question; (3) using meaningful gestures, such as providing an expectant pause, smiling, and short conversational feedback (e.g., "I see", "uh-huh"). The interview was audio-taped and transcribed by the author for further data analysis. Immediately following each interview, the researcher recorded field notes.

4.2.3 Data Analysis

A qualitative research approach known as grounded theory was used to guide the identification of DL evaluation criteria from the interview results. Using grounded theory, "one does not begin with a theory, then prove it. Rather, one begins with an area of study and what is relevant to that area is allowed to emerge" (Strauss & Corbin, 1990, p. 23). Accordingly, grounded theory is useful for inductively deriving a general framework (e.g., a DL evaluation model), when it lacks such a framework, from concrete data collected from social actors' (i.e., DL administrators, developers, and researchers in

this research case) perspectives and experiences. Moreover, grounded theory suggests that such a framework could be discovered, developed, and verified through systematic data collection and analysis.

For data analysis, coding is "the central process by which theories are built from data" (Strauss & Corbin, 1990, p. 57). Qualitative data analysis software, Atlas.ti, was used to develop a coding scheme and assigning meaningful narratives with appropriate codes. The initial coding scheme was developed by incorporating the literature review results (see Table 2.2 through Table 2.7 in the earlier literature review chapter) and the pilot test results, then applying it to the nine interview transcripts as an axial coding procedure. The initial coding scheme was organized into seven categories, with one for DL constructs and the others for the six DL evaluation levels suggested by Saracevic in 2000. Each code in the scheme was named with a combination of DL level and criterion. Hence, a content *accessibility* related narrative was coded under CT-Accessibility whereas service accessibility related texts were assigned with SV-Accessibility.

During the coding process, there might be some new categories not included in the initial scheme. To accommodate this, an open-ended coding technique was applied. After the first coding run was finished, clean-ups were performed to remove less frequently mentioned criteria (i.e., with fewer than three quotes in total) or to merge them to the closest ones if there were any, in light of Auerbach & Silverstein's methodological suggestion (2003). Also, if a code had more than 72 quotes assigned, it was broken into two codes for greater manageability.

To ensure coding consistency across transcripts, a set of rules (see Appendix 4) was developed to guide the coding process. After the first coding run was finished,

iterative coding-recoding reliability checking was executed by the coder (i.e., the author) until two consecutive coding runs for each coding category (i.e., definition, content, technology, interface, service, user and context) reached a higher than 70% consistency rate. The purpose of the reliability check was to ensure an improved validity of coding results. It should be noted that the author as the coder repeated the coding process independently (i.e., without referring to earlier coding results but with the same original coding scheme). Additionally, the second and third recoding processes were carried out only for those categories with less than 70% consistency rate in the previous run. By the end, a maximum of four coding runs were executed before all categories reached the reliability threshold.

After the coding-recoding reliability of over 70% was reached for each coding category of a transcript, another clean-up procedure was executed on individual quotes with inconsistent code assignments. It was found that while a few inconsistencies were caused by error assignment and easily corrected, the majority had to be reexamined and to be carefully reevaluated via revisiting the coding rules (see Appendix 4) and the coding scheme (see Appendix 5). For instance, "It should be there when I need it" with the technology evaluation context was mistakenly assigned with SV-Accessibility during the recoding procedure. As such the correct TN-Accessibility was finally assigned without doubt. However, for the quote of "I want to find articles on pollutions in Raritan [river]. I found a little bit, but I have to go a lot of other places. I don't know" was originally assigned with CT-Comprehensiveness. Although the quote could be interpreted as a DL failure to cover "everything that is within a predetermined scope with respect to subject…" Following the coding rule 2 in Appendix 4, the better code would be CT-

Adequacy, which has been defined as "the extent to which a given DL may provide equal to or sufficient information for a specific requirement." Comparatively, *adequacy* is able to catch the contextual meanings (i.e., my search task for finding "articles on pollutions in Raritan [river]") in the text, whereas *comprehensiveness* is not.

Then, the author examined frequency distribution patterns of all codes (criteria) within and among the six DL evaluation levels as well as among the three stakeholder groups. Meanwhile, a comparison between the code frequencies and the corresponding card-sorting results was made to examine internal reliability among individual interviewees for each interview transcript.

The data analysis results were sent back to the interviewees in early 2006 for "member checking" in order to secure the validity of the study. The criteria in the final coding list would be included in the succeeding survey questionnaire for further confirmation by a more diverse group and larger number of DL stakeholders.

4.3 Online Survey – The Confirmation Stage

4.3.1 Survey Participants

Five groups of stakeholders participated in the online survey: researchers, developers, administrators, librarians and general users. The stakeholders' role identification was based on self-reporting from the demographic section in the survey (see Appendix 6 for the online survey). Table 4.1 outlines the sampling frames and recruitment strategies for the five stakeholder groups.

Table 4.1: Sampling Frames and Recruitment Strategies for the Survey Participants

Survey Participant	Sampling Frame	Sampling Strategies
Administrator Developer Librarian Researcher	Various academic and professional listservs: • ASIS_L (American Society for Information Science & Technology) • jSEES (Association for Library and Information Science Education) • ACRL_Forum • LITA_L, LAMA_WOMAD, LIBADMIN_L, (American Library Assocation) • IFLA_L, IFLA_IT (International Federation of Library Associations) • EastLib (Council on East Asian Libraries) • DIGILIB_L • Web4Lib_L • RUL_Faculty (Rutgers University Libararies) • Other individual DL researchers in LIS, computer, and other related domains	Send the brief solicitation letter to the listserv and individual emails identified from the Web or their publications. The letter contains a hyperlink to the survey URL
General user	Faculty members and students from selected universities in the U.S. that have LIS programs and/or active DL developments	Send the solicitation letter to selected individual email account (e.g. professors and department secretaries at the university selected) and the listserv as the aforementioned and ask them to distribute it to their class/ department/ school listserv

The rationale for surveying faculty members and students from the universities that have LIS programs and/or active DL developments was that they tend to have more opportunity to use and become familiar with DLs, and thus have more valid perspectives on the importance of evaluation criteria. These sampling frames were merely used to identify and recruit various stakeholders. The final stakeholder affiliation in the data analysis was determined by the survey participants' self-reporting in the questionnaire. Three strategies are used to increase response rate: (1) pre-registering these list-serves, being active in these communities so as to increase the familiarity of the author to each community; (2) emphasizing the significance of the survey participants' opinions on the research; and (3) utilizing a drawing for digital devices and thank-you gifts as incentives

to encourage participation. Appendix 5 is the solicitation letter for survey participant recruitment.

4.3.2 Data Collection

During April to May 2006, the online survey form (see Appendix 6) was used to collect the survey participants' perceptions on important DL evaluation criteria, as well as the survey participants' demographics, including age, gender, education, self-reported role in DL innovation, and their DL use experience. The questionnaire was divided into seven sections, with one for demographics, and the other six for the importance ratings on the criteria identified from the exploration stage with either high or least importance perceptions by the interviewees. Each importance-rating section corresponded to a DL level as described by Saracevic (2000). A set of java.scripts were embedded in the survey html file to detect any missing data. When a criterion was not given importance rating, an alert window would pop up, informing the survey participant to complete this in order to continue to the next page.

At the end of each section, an open-ended question was used to collect additional criteria from the survey participants. Survey participants' responses were logged on the server of the author's school at Rutgers University, and then exported into a Microsoft Excel file. The survey participants' names were collected only when they wanted to receive the thank-you gift and/or participate in the drawing for the laptop. However, the names would not appear in data analysis or the report. The author was responsible to maintain confidentiality. The survey protocol was pilot tested by two general users and one DL researcher in early 2006, and approved by the Rutgers University IRB (Institution

Review Board). The pilot test provided useful information for revising criteria definitions so that general users would easily understand them.

4.3.3 Data Analysis

Statistical software, SPSS, was used to analyze the data. For the 7-Point Likert scale in the questionnaire form, "insignificant at all" was coded as 1 whereas "extremely significant" was coded as 7. The distance between any two contiguous points in the scale was indicated as equal. Therefore, the data collected can be regarded as interval, and thus mean and standard deviation were compared for a list of important criteria. Additionally, the ANOVA test was conducted to examine inter-group divergence in criteria importance perceptions. Wherever the inter-group divergence was discovered, the post-hoc technique was employed to further identify the groups contributing to the divergence. In addition to the quantitative data, the content analysis on qualitative data from open-ended questions as for other significant data was also conducted. Meanwhile, their association with the data from the Likert scale was addressed.

4.4 Evaluation Test – The Verification Stage

4.4.1 Digital Library System

The validity of the model constructed was tested through a real situation of DL use. The Rutgers University Library (RUL) Web site (http://www.libraries.rugers.edu/) was the operational DL system for testing. One may refer to the beginning of Chapter 4 for the detailed rationale why the university library Web site has been selected as the DL for the holistic model verification. In general, there were three reasons for the system selection. First, the RUL Web site, as a Web presence of library resources and services, conforms with the definition of a DL proposed by the Digital Library Federation (see

Waters, 1998) in terms of having provided a well defined *user community* (i.e., Rutgers University faculty, staff, students, and the state residents) with a clearly stated *organizational* mission, as well as a presentation of the *organizational structures* and resources. Additionally, the *readily and economically available resources* are *selected*, *organized and integrated*, as well as *persisted* by *specialized staff*. From the site, target users not only can search for physical library *collections*, but also *access digital works (*e.g., online full texts*).* Meanwhile, they may readily seek online *intellectual* assistance from *specialized librarians*.

Another reason for the selection is the ease of getting experiment participants representing various and diverse stakeholder groups. Since the author has been working in university libraries for several years and has developed a good network with those of other than general user groups (e.g., DL developers, administrators, and librarians as special users). As such, the research benefits from input from diverse representative research subjects. Moreover, to the experiment participants, the RUL Web site is presumably a frequently used DL. Therefore, their familiarity to the Web site may become an advantage in the experiment with respect to being able to furnish more experience and knowledge based perspectives on essential criteria for DL evaluation.

4.4.2 Experiment Participants and Their Search Tasks

Various groups of stakeholders who had some familiarity with the RUL Web site were recruited as experiment participants. The groups include general users, researchers, librarians, administrators, and developers with five to six participants in each. Whereas general users were recruited onsite in the two New Brunswick libraries (i.e., Alexander Library for humanities and social science, and Library of Science & Medicine for science

& engineering), the latter four groups of stakeholders were solicited for experiment participation through mailing lists of RUL faculty and staff and individual email communications. As for the onsite recruitment, the author approached potential participants when they came to the library and started to use the library Web site. As for the other four groups of stakeholders, the selection criteria would be: (1) administrators: people whose primary duty was to supervise and manage the library and/or the library Web presence; (2) developers: persons who involved in developing the Web site and/or other RU DL projects listed at http://www.libraries.rutgers.edu/rul/dig_lib_projs/dig_lib_projs.shtml either at a technical or a content level; (3) librarians: reference team members whose primary duty was to use the Web site to help and guide library users in finding information and library collections, and (4) researchers: those who had at least written one published paper about DL(s), and/or had taught at least one course related to DL topics.

These experiment participants were asked beforehand to select a search topic for which they would like to find relevant information through the Rutgers University Library Web site. The rationale for using their own search topics rather than standardized ones was to maximize inter-group divergence with an assumption that different groups might have different interests on the Web site and thereby have different expectations from the site. For instance, whereas a librarian might be interested in finding service information to improve reference, a general user (e.g., student) usually searches the site for his/her coursework. Therefore, it is plausible that perceived importance evaluation criteria vary between the two groups.

4.4.3 Data Collection

These participants' perceptions about the important criteria for RUL Web site evaluation were collected through a post-search questionnaire after they finished searching the site. The data collection focused on the stakeholders' perceptions regarding the importance of each criterion from the holistic DL evaluation model in relation to their search experience with the library Web site during the experiment.

During an evaluation session, right after reading and signing a consent form (Appendix 7), the participants were asked to articulate their information search tasks in a written form. The written form was attached at the end of a short demographic questionnaire (Appendix 8), in which the participants provided general data: age, gender, occupation, self-reported roles as in DL innovation, as well as experience with Web searching and use of the library Web site. Then, the experiment participants started searching the RUL library Web site for their own tasks. While searching, they were encouraged to pay special attention to information, interfaces, as well as various functions on the site rather than those provided by off site commercial databases licensed by the university libraries.

Later, a post-search questionnaire (Appendix 9) was used to collect the data regarding the participants' perceptions of the importance of evaluation criteria at each DL level in relation to their searching experience with the Web site for the task in hand. The questionnaire was divided into three sections: (1) perception of the overall success of the participants' task implementation, as well as their satisfaction with the search experiences and results; (2) perceived importance rating (6-point Likert scale) on the DL evaluation criteria primarily extracted from the proposed model from the confirmation stage; and (3)

open-ended section for perceived significant features on the site that had helped (or hindered) them in the task implementation.

Most criteria in the post-search questionnaire were selected from the proposed DL evaluation model in the confirmation stage. However, a few criteria perceived to be of the least importance by the survey participants were also included. For each criterion, the participants were asked to read a statement about the criterion, and then check off the most appropriate answer in relation to their searching experience with the RUL Web site. A sample statement was "Digital interface should be designed in a way that its essential elements (e.g., color, layout, font, background, terminology use) are **consistent** across sections and pages." In relation to his/her searching experience, a participant could select any of the three options, that is, "not applicable to my case", importance rating on a 6-point scale from the least important to the most important, or "I don't know".

The experiment session with each participant lasted about an hour and took place at public computers in the two New Brunswick libraries for general users and the office computers for administrators, developers, researchers and librarians. Each participant was paid $10 cash for his/her session. The experiment was piloted in July 2006 and finished by the end of October 2006.

4.4.4 Data Analysis

Again, SPSS was used to analyze the distribution patterns of the participants' importance ratings as well as to identify group difference among the stakeholders. The participants' "I don't know" answers were treated as missing data because it had no meaning on the importance of a criterion. Their "not applicable to my case" answers were coded as zero, and were included in the frequency analysis, but not in the descriptive and

inference analyses (ANOVA), because the inclusion could deviate from the mean score and enlarge the standard deviation (SD). Besides, it had totally different meanings from the importance ratings.

The results were compared with the ones from the confirmation stage for the purpose of examining whether the important criteria from the confirmation stage were still perceived to be important when DL stakeholders interact with an operational DL, and similarly, the least important criteria from the online survey were still perceived as trivial in the experiment setting. Additionally, inter-group differences were also examined through the ANOVA test and the results were compared with those from the confirmation stage.

Chapter 5 RESEARCH FINDINGS – THE EXPLORATION STAGE

Since the literature review findings were summarized in Table 2.2 through Table 2.7 in Chapter 2, the section only includes the interview findings. Essentially, qualitative analyses on the nine interview transcripts from the three DL stakeholder groups (i.e., administrators, developers, and researchers) yielded interesting findings, including: (1) sets of consistently perceived important and non-important criteria for evaluations at various DL levels among various stakeholder groups during their card-sorting and open-ended question answering; (2) lists of promising new criteria that were not found in the literature review; (3) essential major consensus and minor divergence among the three stakeholder groups on criteria importance ratings; and (4) feasibility indication of evaluating DLs at the six levels proposed by Saracevic in 2000 but also with three new constructs that were ignored in the existing research body, including activity, architecture, and people (diverse stakeholders). The set of consistently perceived important and non-important criteria served as basic elements in the online survey questionnaire in the succeeding research stage, the confirmation stage.

5.1 DL Stakeholder Interviewees

In total, nine DL stakeholders were interviewed, including (1) three DL administrators (IA1, IA2, and IA3) whose overall responsibilities were to transform RUL from traditional to digital either in general or specific (i.e,. DL system or service) dimensions; (2) three DL developers (ID1, ID2, and ID3) who participated various DL projects at RUL with different capacities; and (3) three DL researchers (IR1, IR2, and IR3) from the School of Communication, Information and Library Studies at Rutgers

University, who had either published paper(s) in the DL research domain, and/or had taught DL course(s) at RU or other institutions.

5.2 Distribution of Criteria Verbalization Among the Six DL Levels

On average, there were 213 criteria-related quotes for each DL level. In Figure 5.1 below, the content and context were the two levels that received above-average number of verbalized perspectives regarding what should be evaluated with 271 (21.25%) and 288 (22.59%) out of 1275 total number of quotes respectively. In contrast, the service level evaluation received the lowest number (116, 9%) of quotes.

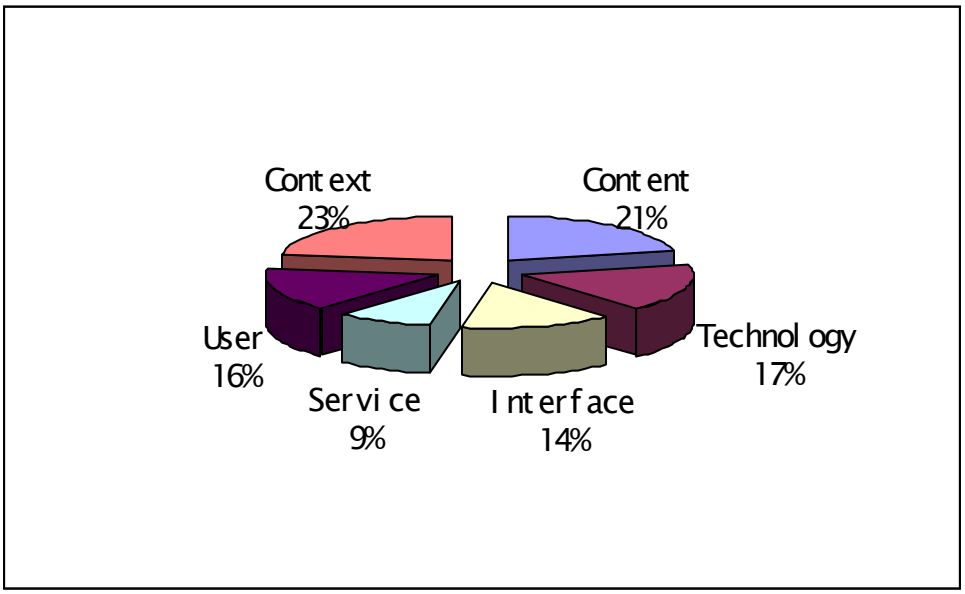

Figure 5.1: Distribution of Criteria Verbalization among DL Levels

Interestingly, among the quotes for various DL constructs, service also received the least number of quotes (33), while content and context are still the two levels with the most quotes identified (117 for content and 120 for context). These results might be associated with the fact that both content and context have the most objects of evaluation (e.g., meta-information, information, digital object, and collection for the content, as well as economic, educational, social, cultural impacts for the context).

5.3 The Most and Least Important DL Evaluation Criteria

Table 5.1 lists the top important and non-important DL evaluation criteria from the open-ended question/answer (QA) as well as the cards-sorting (CS) procedure during the interview based upon the frequency of a given criterion being mentioned (the first numbers in the parentheses), the number of interviewees who mentioned the criterion (the second numbers in the parentheses), or the average ranking order among the nine interviewees for CS. The data are grouped into the six DL aspects. Considering the total amount of criteria at some levels (e.g., service) is less than eight, the table shows the top five important criteria but only three for the least important criteria for each DL level.

It should be restated that the criteria displayed on sorting cards for each DL level were pre-selected from the literature review findings (see Table 2.2 through Table 2.7 in Chapter 2) based upon the frequency of adoption/recommendation. The number of CS criteria for each level was limited to 8 to 11. In contrast, there was no such pre-selection and restriction for QA criteria. What criteria and how frequently they were mentioned was open to the interviewees while they were answering questions, such as "…If you were asked to evaluate digital content, including digital object, information, meta-information & collection, what criteria would you use?"

Table 5.1: Top important and Non-important Evaluation Criteria

	Important Criteria		Non-important Criteria	
	QA	CS	QA	CS
Content	Usefulness to users (32; 9) **Accessibility (32; 7)*** **Integrity (24; 6)** Comprehensiveness (22; 6) Ease of understanding (20; 7)	Usefulness to users (3.7) Accuracy (3.8) Appropriateness (4.1) Fidelity (5.7) Ease of understanding (6.0)	Adequacy (3; 2) Conciseness (5; 3) Size (5; 3) Informativeness (5; 3)	Conciseness (8.4) Scalability (7.9) Authority (7.3)
Technology	Interoperability (36;8) Effectiveness (33; 8) Reliability (27; 7) **Ease of use (17; 6)** Efficiency (15; 8)	Reliability (3.2) Flexibility (3.9) Appropriateness (4.1) Interoperability (5.0) Effectiveness (5.8)	Appropriateness (5; 3) Display quality (5; 4) Security (6; 3)	Cost (7.4) Display quality (7.2) Security (6.2)
Interface	Ease of use/learn (41; 9) Personalizability (20; 7) Effectiveness (20; 6) Appropriateness (16; 9) Supportiveness of HCI** (15; 6)	Ease of use/learn (1.8) Appropriateness (2.3) Effectiveness (3.7) Consistency (5.3) Effort needed (5.6)	Free of distraction (3; 2) Mimicry of real world (7; 2) Aesthetic attractiveness (8; 6)	Personalizability (8.8) Supportiveness of HCI (7.3) Aesthetic attractiveness (7.2)
Service	**Integration (29; 8)** Accessibility (23; 7) **Usefulness (16; 8)** Responsiveness (11; 5) Gaps (8; 5)	Responsiveness (2.3) Reliability (2.8) Accessibility (3.2) Gaps (4.6) Courtesy (6.8)	Cost-benefit (4; 3) Courtesy (5; 3) Reliability (5; 4)	Empathy (8.0)** User's feedback (7.1) Courtesy (6.8)
User	Use/reuse (51; 8) Learning effects (45; 7) Successfulness (17; 8) **Behavior change (17; 5)** Productivity (16; 7)	Productivity (2.7) Successfulness (2.8) Learning effects (3.4) Efficiency (4.7) Information literacy (5.1)	Absence of frustration (3; 3) Immersion (4; 2) Acceptance (5; 2)	Use/reuse (6.0) Acceptance (6.0) Satisfaction (5.3)
Context	Integrity (43; 9) **Managerial support (43; 8)** **Extended social impact (41; 7)** **Collaboration (30; 6)** Sustainability (22; 6)	Productivity (2.3) Outcome (2.8) Sustainability (4.0) Integrity (4.2) Copyright abidance (5.1)	Network effect (6; 3) Outcome (6; 4) Productivity (9; 6)	Network effect (6.6) Compatibility (5.8) Organizational accessibility (5.2)

* the criteria in bold are newly identified from the interview as opposed to the literature review.

** *supportiveness of HCI* was termed as *supportive of H-C interaction* in the instruments, including the interview cards, the survey forms, and the post-search questionnaires in the experiment.

Furthermore, for a given DL level, the card sorting always followed the open question answering. As such, it is not surprising that criteria that were heavily mentioned by an interviewee might not be found in the sorting cards. Similarly, a criterion that was highly ranked by an interviewee might not even be mentioned by the interviewees during the open QA. The transcripts revealed that some important criteria were excluded in the

open question answering due to oversight. For instance, after being presented with the sorting cards of the technology level evaluation criteria, IR3 said, "Reliability, I should have thought about that. Security, that's more important. I guess, I did forget the security matters, and so on." In addition to the recall effect, the variation in total number of criteria between CS and QA as well as the inclusion of new criteria in QA might also have caused the difference.

Therefore, it would be more meaningful to look at shared criteria between QA and CS rather than to look for differences, although a potential reason for a couple of extremes (e.g., *personalization* for the interface, use/reuse for the user level and *productivity of community members* for the context evaluation) might be worth examining. Meanwhile, considering the primary research objective, which is to identify what criteria should be used for DL evaluation, this section will focus on the important criteria perceived by the DL stakeholders rather than the unimportant criteria. In general, over half of the important criteria --16 out of 30-- appeared in both the QA and CS top five rankings (see Table 5.1). The following sections, divided by the six DL levels, will demonstrate these consistently rated criteria with the interviewees' comments.

5.3.1 Content Level Evaluation Criteria

To examine how well a digital collection is developed, digital objects are selected and created, and how well digital information is organized and presented, all nine interviewees agreed that *usefulness to potential users* was the most important criterion. "Then, the *usefulness* goes first, only the user gets to decide that," suggested ID1. Additionally, *usefulness* was considered to be associated with the user's information needs. IA1 commented "…I would say that (usefulness) is certainly important. But one,

it's subjective area, I think. In that, I would have, it depends on who your users are." Further, a successful DL should be able to meet various users' needs, just like what IA2 said—"…I would ask how well does it fulfill its mission for different audiences."

In addition to *usefulness to target users*, *ease of understanding* (i.e., "is it understandable when you are showing people?" as pinpointed by IR3) was also perceived as "one of the important things to learn" (IA2) for DL content evaluation. Similarly, IA1 commented, "it (digital information) needs be clear so it can be understood." Likely, *ease of understanding* should be considered as a premise for making digital content useful to target users.

In contrast to the highly perceived important criteria, *conciseness of information* was rarely mentioned in the interview and consequently had the lowest ranking score from the interviewees. IA1 explained "*conciseness,* that's a goal, but I wouldn't give it as critically high, because I think it still can be effective, even though [it is not concise]."

5.3.2 Technology Level Evaluation Criteria

Due to the existing problematic situation in which federated searching across various collections has not yet been widely operationalized, *interoperability/ compatibility with other standards/systems* was highly rated by the interviewees, as comments from several interviewees reveal. For instance, IR1 complained about extra efforts needed to search individual databases listed on the RUL Web site. The researcher didn't think it was a good DL for searching, because "it doesn't have federated searching at all." IA1 also acknowledged this problematic issue. The RUL administrator commented that their "main goal now is, because it's all individual stove pipe types of information, how do we actually connect all those together and actually find the access

portal to all of that? That might be something similar to what students expect from Google, which you can actually search across a broad array of materials in our digital library and find something." Similarly, ID2 regarded "the ability to aggregate lots of resources and search across them" as "one of the unique features" for consideration. In addition to the *interoperability/compatibility with other systems* at the time being, some interviewees also expressed their thoughts about *system compatibility* over time. ID2 predicted, "ten years from now, the software that we're now using is going to be outdated. So, we really need to be able to easily replace pieces."

In addition to the specific goal of searching across different systems and migrating from one system to another, the *system effectiveness* of satisfying various expected purposes was also frequently mentioned, as summarized by ID3 – "[it] can do most of the things we are looking for." Meanwhile, not only should a DL provide technologies supporting various collection and service goals, but the support per se should also be *reliable.* ID2 spoke of the importance of *reliability*: "I learnt when you do a global searching replacement, you should always do a backup first and think once or twice before you do it." Similarly, IA2 regarded the absence of "interruption to collections and services to users" as the most important aspect of digital technology.

Whereas the majority of the interviewees were concerned about a system's trustworthiness to work smoothly, very few focused on digital software/hardware's capability to protect the system as well as the users' personal information. IR2's perspective on the *security* issue was twofold. On one hand, the interviewee thought, "information as public good' was "not so important to me. I don't care." On the other

hand, the same person was concerned about "the government[']s] hanging around to spy [on] our [online] user behavior."

Display quality was another trivial criterion in these interviewee's perspectives. IA3 explained, "I didn't think of [it], because I assumed [it is] going to be excellent." IR3 even thought sometimes it might not be practical if display quality is overemphasized--"Maybe I should say appropriate quality, since occasionally I was getting mad with people who do things like demand sound conversion at frequencies above [the] hearing range of the human ear."

5.3.3 Interface Level Evaluation Criteria

As the most important criterion for interface evaluation, *ease of use* ("intuitiveness" and/or "transparency" in the words of these interviewees) was rated highly by the nine interviewees. The comments from IA2 and IR1 are representative. IA2 commented, "…the ease of use would probably be the top… A really well designed interface should be almost transparent, and should provide the most transparent access to services and information…It should also be intuitive and easy to understand. It should not require a lot of help screens." IR1 further explained, "I expect from the interface that makes transparent to me for doing the task, and I don't have to fight with the interface. I just want the interface to disappear. So, I connect directly with the content, with the searching, and with everything else you know, as if nobody is between me and the content and the search mechanisms."

To do so, one promising strategy is to keep users in mind during the design and implementation process, as suggested by ID3: "Basically, when you design the interface, you have to think in terms of the user, not the developer…" It was considered to be

especially important to make the interface meet various users' needs, and work compatibly with their backgrounds and behaviors. For instance, for RU DL projects, IA1 told the author, "What we're trying to do is to create portals that allow a student portal, [and] a researcher portal. I would say I think that is important…" Similarly, for his students' term project for his DL course, IR3 said, "And the thing I tried to warn them about the most as the greatest failure in many projects is [the] failure to anticipate [that] users don't know anything about the subject." Likely, for interface evaluation, the interviewees agreed on the importance of *appropriateness to target users*.

The primary role of the interface is to assist users in finding the information they need. Therefore, it is not surprising that *effectiveness* was considered by the majority of the interviewees to be one of the important criteria for interface evaluation. *Effectiveness* was defined as the "ability to find the information" by IA1, and "how well I can accomplish tasks with that particular interface" by IR1. Further, according to ID2, not only should an interface be able to bring the desired information up to the front, it also should not bury users with overwhelming results. The interviewee explained, "you have a lot of searches, and each time they are getting thousands [of results], there is something wrong…"

Interestingly, although *aesthetic attractiveness* was one of the lowest rated criteria, some interviewees noted the need to consider *aesthetic attractiveness* in interface design. ID2, the Web developer, mentioned receiving "complaints about colors" from users. This experience implies the need to develop DL aesthetically, because some users care about the aesthetic aspect of a DL. ID3 further suggested to have people with "some art background actually design the interface, so that it is very pleasing to people. So, when

people sit down, the first look they get, they [don't] get bored. So, the interface should be interesting in terms of looking."

5.3.4 Service Level Evaluation Criteria

While digital service was regarded to be essential to a DL, *access* to a specific service without physical or financial barriers was one of the important criteria for interviewees. Specifically, to reach a real service person online was frequently mentioned by the interviewees. "The live reference [person] should be right there," and "it would be wonderful if I could have a chat with a curator, search some unusual materials, and they could give me …" said ID2 and IR2 respectively. Similar to ID2 and IR2's comments, IA3, as an administrator, was hoping, "…library personnel can go where the users are. So, if the users are sitting in front of their computers in their offices, I want our staff to be able to get there with them. So, I think I am thinking about live chat to some extent, live reference…" There were representative voices from the three stakeholder groups advocating for live reference service. Unfortunately, it is hard to find a well-developed service nowadays. IR1 complained, "The digital libraries have the same problem that many of the digital enterprises have. This is a connection with a live person when you need it. And there should be a way with which the digital library provides me with the connection, if I need somebody live."

The lack of virtual reference service is one of the big *gaps between expectation and perception.* There were services that users wanted but for some reason were not accessible, or were of unacceptable quality. For instance, IA1 commented on a faculty member's request for a bibliographic formatting service, "It is interesting… the faculty member wants us to build into the digital library this functionality. It's something that we

would never have thought about. And I think as our users get more accustomed to what we can put [into] it, new services will probably emerge that we haven't thought about." Accordingly, IA2 suggested that DL developers "need to really target users' needs, not what we think the user needs, or more or less what the user needs in a given context."

Responsiveness ("synchrony" in the interviewee's words) was another important criterion highly rated in the interview. Whereas ID2 didn't think email reference was good because "You have to wait for at least for 24 hours," IR1 had a negative experience of using an interlibrary loan service whereby he could not just "request it and then get it within a reasonable amount of time."

Among the highly rated criteria, there was a new criterion worth noting, *integration of digital services into the users' information seeking path (i.e. contextual service).* Nearly all interviewees verbalized their thoughts on how digital services should be integrated into the main flow of DL use in a transparent way so that users could immediately access it when they wanted. IA1 suggested an evaluation of digital reference in terms of "do you make it easy for them, then to immediately "Ask a librarian", or they have to get out of the digital library, go to another little thing and find it," when "someone is searching a database." She regarded "…the integration of the service into the use of a digital library" as one of her "top priorities". IA2 had a similar comment on the contextuality of digital service:

> The second would be the help itself needs to be fully integrated into the primary service. So the agent services that support our digital library collections and services should be well integrated into overall service itself or overall collection itself. So students do not have to step out of the frame of what they are doing to get help...the contextuality of the service is critical. Does it fit in the user's workflow?… Does it integrate seamlessly with overall the collections and services with which they need help?"

IA3 gave a vivid instance of a contextual DL library service that could be provided. The administrator said, "if technically a system can figure out if someone is searching for something and they are trying IRIS (the RUL OPAC) six different ways and it just doesn't show up, then maybe they should be presented with an option, or the request should automatically go to PALSI or an option for an interlibrary loan form, something like that." Emphasizing contextual digital service as well, IR2 analogized "unwanted services" to "information overload." The interviewee regarded Google as an exemplar of an undisrupted service provider, because "they have algorithms to match your search with ads. But the ads are on the side, so it's not intrusive."

Courtesy was perceived as the least important. ID2's explanation is plausible, noting, "courtesy and empathy. It's important, but I thought if a digital library is designed very properly, actually they are not going to interact with staff very much. "

5.3.5 User Level Evaluation Criteria

While acknowledging the value of indirect assessment on DL performance by using individual or groups of DL users as the objects of evaluation, the interviewees regarded *learning effects, successfulness of task completion,* and *productivity changes before and after the DL use* as three most important criteria at user level. The interviewees expressed similar viewpoints when they commented on *learning effects*. Whereas IA2 considered, "the most important criteria would be evidence of adaptive learning…" IA3 wanted to "see them [students] have better grades…" and IR2 regarded enabling "new types of learning" and "educational relevance" a success. IR3 even described an ideal way of measuring the success. The researcher said, "…suppose you are running a class, and you have a choice to either give the class reference materials on

paper or giving them the reference materials in electronic form, will the students learn better in one way over the other?" However, he confessed that this approach might be restricted by the current practice of human subject protection.

Because all nine interviewees came from a university, where education is the core mission, it is not surprising that *learning effect* was perceived as the most important criterion for user level evaluation. Comparatively, s*uccessfulness of task completion* was applicable to various information seeking circumstances in addition to education. Not only was it considered to be important in terms of the actual degree to which an information seeking task is accomplished, but users' perceived *successfulness* was also suggested by IR3. In addition to the immediate output of DL use, *productivity* as a temporal outcome of DL use was also highly regarded by these interviewees. IR3 thought, "productivity change would be the most important." The productivity changes could be reflected in diverse aspects, from "doing a better lecture" (IR1) to "more productive lives in the information society." (IA3)

5.3.6 Context Level Evaluation Criteria

For assessing DL outcomes at the context level, including institutional, social, cultural, economic, legal, etc., all nine interviewees agreed that a successful DL ought to be integrated into the goals and practices of its larger context(s), especially the institution of the DL. *Integrity* was regarded by ID2 as "pretty important." When it was applied to the university library setting, IA3 commented, "One of the first parts is, to me, that our digital effort has to support the mission of Rutgers University. That's primary." *Integrity* was so important in IA2's perspective that it should be seriously considered from the very beginning of DL development. The administrator commented, "the best measure would

be…[to] have an understandable data model, and data and service model behind your initiatives that reflect the goals and needs of your larger organizations and of your users." Further, according to IR3, not only should a DL be smoothly integrated to the local institution, but it also should be well adapted into broader contexts. For instance, the interviewee argued for the necessity of assessing a DL in terms of "whether it conforms [to] the society's norms."

To develop a DL requires support. To maintain a DL needs even more resources. Some interviewees (ID2, IR1, IR3) explicitly pinpointed the importance of DL *sustainability*, while others (IA1, IA3, ID1, IR2) expressed their viewpoints implicitly. IR1 emphasized, "That [sustainability] is of course a very critical issue. It is the most important issue…The permanence of a digital library is a huge criterion." IA3 further explained why the issue is so important, saying, "Oh, we need money. We always need money, because we buy these equipment, people, the ACM digital library, resources, the ability to program…. It's an expensive proposition."

The interviewees viewed *network effects* (i.e., relationship with other network resources) as trivial. They tended not to care much about the number of incoming links to a DL. IA2 explained, " The incoming link is not as good anymore, because everybody knows how Google works." Instead, the interviewee suggested, "a better measure would be how many outgoing links referring to other related and complementary resources." Similarly, IA3 commented, "I think we got to have to find more ways to relate to other collections or other libraries."

5.4 Emerging New Criteria

Some promising DL evaluation criteria (see Table 2.2 to Table 2.7 in Chapter 2 for the one in italic texts), which were not covered by the existing research, have been identified from the interviews. They include:

- *integrity of information*
- *technological ease of use*
- *service integration into information seeking path*
- *user's behavior change*
- *copyright reform/fair use*
- *integration into organizational practice*
- *extended social effect*

Among the six DL levels, the interface evaluation seems to be the most thorough one with only 4 out of 16 (25%) new criteria, whereas half of the 12 criteria (50%) at the context level were not found in the previous studies. The finding supports Saracevic's argument (2000) in terms of lacking existing evaluation studies at the context level. It further implies the necessity and plausibility of assessing DLs at this level as recommended by several researchers (Bishop et al. 2003; Marchionini 2000; Marchionini et al. 2003).

5.4.1 Content Level Evaluation

When being asked to verbalize essential criteria for digital content evaluation, the interviewees frequently mentioned *integrity of information over space/time,* which was defined as the extent to which a DL information/collection can be incorporated with other resources to form a complete unit over time and space. IA1 pinpointed, "A big thing for

us obviously is archiving and preservation of digital libraries, which I think is a major issue, which is how do you assure the integrity of information over time?" Whereas IA1 was concerned about the integrity over time, IA2, ID2 and IR2 highlighted the integrity of the resources. IA2 described the latter as DL's "ability to relate information [to]…each other to give you a richer response to queries." The administrator further commented:

> You can have a little bit of everything. But you don't get in-depth in certain topics, you don't have materials related to each other, so that users can scaffold that information to come up with, to synthesize something for themselves. So, it's not enough to find some little pieces of discrete information. The information has to relate to each other. It has to have some coherence and integrity.

Similarly, ID2, as a DL developer, was concerned about "what do you put into it [a digital library]? What makes it coherent? The developer criticized an on-going DL, "If you look at the…Digital Library, to a certain extent, it is almost random." Moreover, IA2 suggested the integrity should be also reflected in the interrelationship with other DLs and Web-based resources in additional to those within a given DL. As such, instead of emphasizing the increased number of incoming links, IA2 highlighted the necessity of counting "how many outgoing links referring to other related and complementary resources."

Although *uniqueness of collection* didn't receive the same importance ranking as *integrity of information over space/time,* four interviewees from all three stakeholder groups expressed the need to consider it. IA1 thought it would be nice to have unique content by asking a similar question: "is the content something that would be unique contribution to overall information access?" IA3 was proud of some of the digital collections at RU, because they are "special." When talking about the collection of Bill Clinton's 90 million email messages, ID2 asserted its potential value for research. The developer said, "That never existed before. So, these new areas for researchers are really

promising, I think." In grading his students' term projects (i.e., DL development proposals), IR3 weighted heavily the novelty of their proposed DLs. The interviewee told his students:

> I don't want you to duplicate something that is out there now. So, if somebody said, 'I'd like to scan the journal of American Historical Society," I said it was not a good project, because JSTOR has already done it. And then he said, "Well, I want to [digitize] the newsletter of the Montclair Historical Association." I said, "ok," because nobody else had done that.

5.4.2 Technology Level Evaluation

When asked to verbalize their thoughts about what criteria should be used for digital technology evaluation, the majority mentioned *ease of use*. Likely, the notion of *ease of use* was twofold. One is *ease of use* for end users of a given DL, and another is *ease of use* for people who develop and manage the DL. For instance, IA3, on the one hand, "wouldn't want to take tons of people and tons of time to manage [a digital library]." Instead, she wanted it "to be as fairly straightforward as possible." On the other hand, she wanted digital technology "to be versatile, in that users wouldn't do much…They [users] wouldn't have to download files in order to use something." ID2, IR1, IR2, and IR3 also noted *ease of use* on the user end. ID2 commented, "The digital library is going to grow. The issue is making special software really transparent to users. I really should not have to worry about plug-ins. People don't like to do that even though it is very easy to just hit the link." Whereas IR1 and IR2 required "simplicity" in digital technology, IR3 straightforwardly suggested digital technology should be "ease of use by the final user…low demand for expertise." Another instance of *ease of use* for DL development and management was pointed out by ID3. The DL developer mentioned *ease of use* several times when he was talking about the reasons why their DL projects

adopt one management system supporting a particular program language (JAVA) as opposed to others.

5.4.3 Interface Level Evaluation

While important criteria for digital interface evaluation showed little divergence between the interview findings and the existing research, one emerging criterion is worth noting. When summarizing the advantages of a DL as opposed to a traditional library, ID2 thought, "there are what I called vertical features. Once you have something in digital format, if this is an artifact, you could rotate it and examine [it] from all sides, and zoom in. These kinds of things, sometimes, are more difficult and impossible to do in a more traditional library." Similarly, the degree to which a DL can mirror reality, that is, *mimicry of the real world,* was perceived by IA2 as a promising aspect to be pursued. The administrator said, "Another thing, I think, that can do very well, and people try very hard to do, is [that] it can bring a lot of complex things that are real, a lot of complex, real world phenomenon. You can select and choose among them to present [a] mosaic of reality that is easier to process than being in a real world."

5.4.4 Service Level Evaluation

Integrity to information seeking path was the most frequently mentioned service evaluation criterion during the interview. Eight out of the nine interviewees suggested a digital service should be well integrated into the main flow of DL use so transparently that it can be immediately reached by the users whenever and wherever they want. IA2 commented, "ideally [how] we really want to evaluate it is where does…the service…actually fit in the life-long or the daily use of users…the contextuality of the service is critical. Does it fit in the user's workflow, and get their job done, as directly

responsive to what they need, and neither too much nor too little?" The preceding important criteria section (5.1.3.4) has more verbalized thoughts on this criterion from other interviewees.

5.4.5 User Level Evaluation

As mentioned earlier, compared to those for a narrower sense of DL evaluation (i.e., technology, interface and service), more new criteria were found for DL evaluation at the user and context levels. These new criteria go beyond the coverage of traditional information retrieval system evaluation. *Behavior/workflow change* is one where DL performance is indirectly assessed through indicators of its users instead of the DL per se. After responding to the assessment approach by saying "that's a very good way of evaluating user behavior after they use [digital libraries]," ID3 suggested an approach to user measurement in terms of "how this affects their way of working. Do they still come to the library, or do they just stay at home or stay in a computer room and get everything from the computer?" Speaking from her cultural analyses of a number of national DL innovation projects in Europe and North America, IR2 confirmed that the DLs have "engaged people in new dialogue of materials…it's kind of document that can provide new interactions for new types of users." According to the researcher, the changes represented "success". Meanwhile, ID2 observed, "[people] are spending a lot of money on printing" in a DL environment. Additionally, the developer voiced his concerns about the negative changes in how people read a book in a digital environment as opposed to in a traditional library setting. Immediately after claiming "people are going to read fewer and fewer books from beginning to end," ID2 further commented, "I don't think this is

good. People need to read, and they need to read the whole book, not just a few paragraphs out of each chapter."

In addition to a consensus on the importance of evaluating users' *behavior changes*, several interviewees suggested that one might assess DL outcomes by examining the extent to which users show their understanding and support of a DL. The *supportiveness* of DL users could be measured, as suggested by IA2, by looking at whether a DL "is handed off from one user to another. Like users as educators, they used resources, and passed [the resources] on to their students. Or students used resources and passed them on to other students in the project or in their presentation." Whereas IA2 was thinking about the support from general users, IA3, from another administrative angle, highlighted the beneficial effect of getting support from some "influential people…people who can explain your story throughout the university." Meanwhile, IA3 also pinpointed the need for general DL users to understand the huge amount of investment needed to build a DL. The administrator hoped that DL users might recognize "all of these [digital resources] are expensive and it takes effort", and thereafter "In the long run, a great big benefactor who used us might come to us and say the library is great, and I am going to give a million dollars to the library in order to make more information for better students for better grades. So, I think that the whole cycle of that…the support, it is just an understanding of that it is effort and that is valid."

5.4.6 Context Level Evaluation

Six new criteria were identified from the interviews for evaluating how well a given DL fits into the broader organizational, social, cultural, legal, and economic

practices, and what impacts and effects the DL may have on these contexts. These criteria are:

- *integration with organizational/social practice*
- *managerial support, extended social impact*
- *collaboration/sharing, copyright reform/fair use*
- *scholarly communication.*

Integrity with organizational/social practice, as the criterion perceived as the most important, was reported in the preceding section of the most and least important DL evaluation criteria (5.3.6), and therefore will not be repeated here.

Whereas the existing criterion, *sustainability*/*affordability*, is usually associated with financial factors, *managerial support,* a newly identified criterion, refers to the extent to which DL development/maintenance is supported by human or physical resources. IR1 suggested, "You can also ask questions: how well is it managed the whole thing?" When talking about the rationale behind one of the library DL projects, IA1 said, "The reason we went for it is that we felt that there are lots of materials within New Jersey in small libraries. They don't have the local expertise or financial resources to be able to digitize the materials…" Obviously, rich collections and local expertise are the two dominant supporting factors for the DL. According to IA1, when *managerial support* is limited, the library administrators have to "come to some consensus about priorities in developing digital libraries." IA3 also verbalized her experience with human resource management, describing it as "very big…to provide [the] right training to [the] right people at [the] right time." Additionally, the interviewee provided an example of how Web masters and programmers are crucial to DL development. In addition to *managerial*

support within an institution, sometimes outside support is crucial as well. For instance, with respect to standardization in DL developments, IA2 pinpointed the necessity of having "some communities, like the Digital Library Federation, NISO, and others, that are continually reassessing the standards out there…"

Similar to the administrators, the developers also had similar perceptions of *managerial support*. ID1 described his experience about the need for taking down a given Web-based subject guide from the library Web site, because the content maintainer, "the librarian and that position doesn't get filled," and the subject guide was not able to be updated. ID2, from the DL architecture aspect, argued for the need to collect administrative data, such as circulation statistics. He thought, "those kinds of statistics have to be part of a digital library, and really part of the administrative function. It's got to be examined…But you have to [be] able to manage." DL management should start even before actual development, ID3 further commented. Although admitting a few negative effects due to potential cultural differences between collaborative units, the developer judged an ongoing DL project as being successful because it "is well defined from the beginning," "the [development] team has very good leadership," and "the whole team… is a very good team."

Being efficient, effective and also valuable to DL innovation, *collaboration/sharing* was highly recognized by the interviewees. This is the extent to which DL innovation processes and products are shared among stakeholders and/or institutions. IR2 found "a trans-institutional corporation basis even from the beginnings" from her cultural analyses of several national DL projects. Such a trans-institutional collaboration was perceived to be extremely important. According to IA1, without

resource sharing, NJ Digital Highway, a large collaborative project between RUL and other libraries in NJ, would never have been accomplished, because "we (RUL) don't have all the content that the other people do…But they would never be able to develop the infrastructure on their own." Furthermore, the administrator predicted "the whole idea of collaboration in digital libraries is going to continue to elevate, and I think, there are probably fewer places taking on a complete role in developing digital libraries from A to Z." Sometimes, institutional contribution to a collaborative DL project might not be so equal. According to IA2, RUL, as the largest academic library in New Jersey, expects "to serve as a platform for individual libraries and archives that cannot build their own cultural heritage initiatives. So we then can provide a local focus, and a local presence that they can attach to their own Web sites." In her view, there should be an institution that takes a leading role to help other organizations with insufficient resources through resource/technology sharing.

Not only could *collaboration/sharing* take place among institutions, it also could be applied within an individual institution. IA3 and ID3 mentioned collaborations among different departments and people within the RUL setting. Moreover, in addition to collaborative DL development, *collaboration/sharing* also could be encouraged for DL applications, such as research and learning. ID2 perceived "an opportunity through a digital library [where] you can join others [who] are doing research on the same thing."

Acknowledging the need for copyright protection, the majority of the interviewees expressed a demand for *copyright reform/fair use,* which challenges the notion of *copyright compliance*. IA1 described an ideal situation within which "the whole copyright issue is being able to use the information in ways you want to, and not be

confined by copyright restrictions." The administrator further noted "a great need in copyright legislation to understand that it's important to be able to build on the work of others. That's how progress is made." Similarly, ID3 would "like to see there is a movement for information to be shared. So, that's not just owned by …[a] few people or a company, but rather being viewed as a global good, public good for the world. So, that's a big understanding of information value." Otherwise, ID1 warned, "We are going to wind up paying for every single use. And the concept of fair use may be lost in that." Fortunately, according to ID2, the issue is getting resolved, saying, "some publishers are figuring out they've got to either have their authors retain their copyright, or alternatively it's a limited time, let's say six months."

There were various *extended social impacts* of DL innovation described by the interviewees. They include: (1) facilitating multi-disciplinary research and other research activities (IA1, IA3, ID2, IR2); (2) improving social/economic status (ID2, IR2, IR3); preserving knowledge and culture (IA2, IR2, IR3); (3) advancing democracy and legislation (ID3, IR2, IR2); and (4) contributing to the national and/or international society (IA1, IA2, IA3).

Overall, the higher the DL evaluation level (i.e., context and user), the more new criteria are identified from the interviews as compared to those for lower level DL evaluation (i.e., technology, interface, and service). As a result, the findings might fill the gaps in the existing research whereby few studies and criteria were found for context evaluation.

5.5 Consensus/Divergence Among the Stakeholder Groups

In addition to the variances in the importance of DL evaluation criteria ratings within and/or among DL levels in general, some consensus and divergence are also identified among the three stakeholder groups (i.e., administrators, developers, researchers) regarding what are the important criteria for a given level DL evaluation.

Table 5.2 lists the consensus and divergence criteria from the card sorting results. The reason for using card sorting rather than open question/answer results was that the criteria are identical among interviewees in the card sorting and thus readily for comparing the interviewees' perspectives. In contrast, there was too much variance in the open question/answer regarding what criteria were mentioned by the interviewees. The criteria consistency in the former makes the comparison among stakeholder groups feasible. The choice of consensus or divergence was based upon comparing the sum of the ranking value from each group for a given criterion. If there was any group value larger or smaller by a factor of 2 for any of the other two groups, then this criterion was considered as having a divergent ranking among the stakeholder groups. Otherwise, it would be considered consensus. For example, the sum of the ranking value in the card-sorting for content *usefulness to target users* is 15, 5, and 13 respectively for the administrator, developer and research groups. Therefore, the criterion is considered to be much more highly ranked by the developer group than the other two, and thus is categorized as divergent. In contrast, content *appropriateness for target audience* has 11, 12, and 14 as the sum of ranking values for the administrator, developer and research groups. Accordingly, it is considered as a criterion with agreed-upon importance among the three stakeholder groups.

Moreover, the criteria with higher importance rankings (e.g., *usefulness of information, technological reliability, interface effectiveness*) tend to have more divergence and less consensus than the lower ranked criteria (e.g., *conciseness, security, personalization*). Also, the service level and the user level diverged less (one out of the top five) while the other four levels had two or more perceived important criteria with wide variance.

Table 5.2: Consensus/Divergence Criteria Among the Interviewees

	Consensus	Divergence
Content	**Appropriateness for target audience*; Fidelity; Ease of understanding**; Informativeness; Authority; Scalability; Conciseness of information	**Usefulness to users**; **Accuracy;** Comprehensiveness of collection; Timeliness (freshness)
Technology	**Flexibility; Appropriateness for digital information; Interoperability / compatibility**; Efficiency; Security; Cost	**Reliability**; **Effectiveness**; Comfort for use; Display quality
Interface	**Ease of use/learn; Consistency**; **Effort needed;** Efficiency; Error detection and handling; Aesthetic attractiveness; Supportiveness of HCI ; Personalizability	**Appropriateness to target users; Effectiveness (e.g. precision/recall)**
Service	**Responsiveness; Reliability; Gaps between expectation and perception; Cost-benefit;** Use/reuse; Courtesy; Positive feedback/reaction; Empathy	**Accessibility**
User	**Productivity; Learning effects; Time of task completion; Information literacy**; Satisfaction; Acceptance; Use/reuse	**Successfulness of task completion**
Context	**Affordability/ Sustainability; Integrity into organizational practices; Copyright compliance;** Organizational accessibility; Compatibility; Network effect	**Productivity of community members; Outcome against predetermined institutional goals**

* The criteria in bold text are the ones within the top five-importance list

Compared with the consensus criteria, the divergent criteria merit more attention. In particular, it is worthwhile to find out how and why the divergence exists. The following explores the two issues by looking at ranking ratings among groups and individuals as well as interpreting the interviewees' input.

- Usefulness of content to target users

While all three developers unanimously gave *usefulness to target users* the top ranking, there was disagreement in the other two groups on its importance in digital

content evaluation. For instance, IA1 and IR2 ranked it 10th among the 11 criteria where the other peers gave it the top 1 to top 3 ranking. In IR2's perspective, even "if the [contents] are only relevant to a very limited number of scholars, [they are] still important relevance." IA1 regarded usefulness as being very subjective and varying among users. The administrator said, "[usefulness] is certainly important. But one, it's a subjective area, I think. In that…it depends on who your users are."

- Comprehensiveness of collection

Similarly, IA1 thought *comprehensiveness* should also be well defined. The interviewee commented, "I don't think any collection has to be comprehensive. But comprehensiveness might be based on what you are saying this is…So, I would say, it's important, it's critical, but it also has to be defined. It isn't really a global thing." Being different from *usefulness to target users* whereby the divergence was due to sharp disagreement within a single stakeholder group, the cause for the divergence in perceptions of c*omprehensiveness* is more likely to be associated with group differences. Specifically, whereas all researchers agreed on the importance of *comprehensiveness*, all administrators and all developers assigned the medium or least importance rating to the criterion with an only exception-- ID regarded it as being "very important".

- Accuracy of information

In terms of whether digital information needs to be accurate, all administrators unanimously considered this a critical issue, whereas the developers gave it moderate consideration. Meanwhile, IR1 deviated from the other two researchers and assigned a lower ranking score to *accuracy*. Unfortunately, the underlying reason for the score was not given during the interview.

- Timeliness of information

Although the majority of the interviewees considered *timeliness* as being comparatively unimportant, there were three interviewees with two from the administrator group and one from the developer group who ranked the criterion highly. IA1's comment is representative: "…it's absolutely critical that whatever included is accurate. I think that is the most important. And it has to be relied on, so it's got to be what it is. And it's got to be up to date. "

- Reliability of technology

While all the administrators and developers provided top rankings to technology *reliability*, two out of the three researchers (IR1 & IR3) gave it a moderate score. Unfortunately, no detailed reason was given.

- Effectiveness of technology

To the extent to which digital hardware/software has to be developed to achieve pre-determined goals, the administrators ranked *effectiveness* of technology as a moderate issue, and the researchers thought it was not so important. In contrast, two developers (ID1 & ID2) gave it the highest ranking. Perhaps the reason why ID1 assigned the highest score is that effectiveness "will work for all equipment," whereas other digital technology evaluation criteria (e.g., *security*) were considered to be appropriate only for certain technologies. In contrast, IR2 thought digital technology did not have to be restricted to achieve a "pre-determined goal." Instead, sometimes it could be useful if one expects to find texts on a given topic and also get images, and audio on that topic. The salient effect was considered to be more important instead.

- Comfort of use

Interestingly, for unknown reasons, two interviewees from the research group (IR1 and IR2) assigned the highest score to the degree to which a digital device can be used comfortably. In contrast, all developers and administrators as well as IR2 thought this was a trivial issue.

- Display quality

The same two researchers (IR1 and IR2) thought it was important for digital technology to display information at a high standard while all remaining interviewees, including all administrators and developers, varied in their perceptions. One plausible reason is that they think this issue was taken-for-granted and needed no attention. For instance, IA3 commented, "I am looking at display quality. I didn't think of this thing, because I assume [it is] going to be excellent." Another reason is from IR3, who thought "appropriate quality" might be more critical than over-emphasizing the quality issue. The researcher shared from his own experience, "since occasionally I was getting mad with people who do things like demand sound conversion at the frequencies above [the] hearing range of the human ear. [It] should be an appropriate quality." Likely, the two researchers (IR1 and IR2) might have a shared personal preference for high-end technologies. This implies that individualized preference may have an impact on the importance rating of *display quality*.

- Interface appropriateness to target users

While this issue received the highest scores from almost every interviewee in the study, only ID2 assigned a lightly lower score to *appropriateness to target users* than the others. The developer perceived that as long as a digital library is on the Web, it should be for everybody. And "in the World Wide Web, as I said, it's really hard to get in touch

with your users, because they are anybody."

- Effectiveness of interface

Interface *effectiveness* is defined as the extent to which the interface is capable of helping users in achieving their goals and objectives, in particular, finding information for given information seeking tasks. Conventionally used measures would be precision and recall, which are based upon the relevance of documents retrieved from the information seeking tasks. A comparison of the scores among the three stakeholder groups shows that whereas all developers thought it very important, two from the research group (IR1 and IR2) gave *effectiveness* a low score. IR2 argued:

> It's difficult to get it in a way that meets the relevance of a particular research project... Maybe it is because the nature of searching is too vague. You have to go through certain protocols. So, I do think there is an ideal searcher, a librarian, who can retrieve, who can be a mediator, to use fully those components of a digital library. I don't know if it's a really politically sound way to think about them.

Considering that it is impractical to emphasize interface *effectiveness*, IR2 suggested, "Just standard cost-effective use is ideal."

- Service accessibility

The divergence among stakeholder groups on *service accessibility* is not strong. Only two developers (ID1 and ID2) and one researcher (IR2) provided a moderate score, while the others, especially all three administrators, regarded it as a very important issue. Even for the three moderate score givers, their words still demonstrate a fairly important perception. For instance, IR2 expressed the wish to see a live chat service staffed with librarians. The researcher commented, "it would be wonderful if I could have a chat with curator, search some unusual materials,…So, in that sense, if there is a human out there, a human voice, [or] human presence out there, that I could tap into it, that would be lovely. This would be [the] ideal situation." ID2 had a similar hope. The developer said, "Yah,

the live reference should be right there."

- Successfulness of task completion

For DL evaluation at the user level, the researchers assigned the most important score to *successfulness of task completion*, but two from the administrator group (IA2 and IA3) thought the importance was moderate. However, the underlying reasons for this difference were unclear. The only plausible one based upon the author's personal reasoning is that in accordance with the university libraries' core mission (i.e., to advance teaching and learning), the administrators tended to rank *learning effect* and *information literacy* much higher than other criteria at the user level. Presumably, this is why the administrators ranked *successfulness of task completion* lower.

- Productivity of community members

Interestingly, all researchers ranked this criterion at the very top, but the administrator group rated it as less important: two (IA1 and IA3) ranked it 3rd and one (IA2) ranked it 6th. However, in general, except for IA2, the rankings by the other interviewees were all within the top three. Hence, the group divergence is vague.

- Outcome against predetermined institutional goals

Similarly, except IR2, all interviewees ranked very highly *outcome against predetermined institutional goals*. Thus, the variance should be considered as individual rather than group associated.

To sum up, the higher a digital level is, the more consensus was reached by the stakeholder groups based upon the comparison of cards-sorting results. It is presumed that the content and technology levels are usually hidden in a DL system and require expertise to fully understand them. For instance, the administrator said, "to tell you the

truth, I can't speak to the hardware issue at all." Similarly, the developer had difficulty speaking about content evaluation. The interviewee admitted, "I can't clearly describe and sort [through] scholarly journals and foreign languages. I don't have opinions." This, from another angle, implies that a given stakeholder group (e.g., administrators) might be appropriate interview participants for judging certain DL aspects (e.g., context) while they might provide unreliable rankings when asked to assess the other aspects (e.g., content, technology etc.).

It was also found that the importance ranking divergence varied between perceived important and unimportant criteria. Essentially, the amount of consensus criteria is greater than the amount of divergent criteria. Moreover, comparatively, more divergent criteria received top importance ranking scores. These findings imply the plausibility of conducting general evaluations satisfying the preferences of various DL stakeholder groups. However, considering the divergence, especially for evaluation at the content and technology levels, tailored evaluation for meeting specific stakeholder groups' needs might also be necessary.

5.6 Comparison on the Interview and the Literature Review Findings

There were 77 criteria in the final coding scheme within which more than half of the criteria--55 (70%)--appeared in both interview and literature review, while 22 (30%) were new to the existing research body (see the section 5.1.4. for details). Meanwhile, only a small portion (7 out of 87, 8%) of the criteria from the literature review were not on the final coding list. They are:

- *information readability*
- *growth rate of a collection*

- *technological complexity in query support*
- *accessibility of interface features (e.g. additional help link),*
- *error detection/ handling, users' preference*
- *difference before and after a service intervention*

Table 5.3 summarizes the accumulated criteria from the earlier literature review, the interviews, as well as the combination of the two for the six DL evaluation levels. The table shows that among the six levels, the interface evaluation seems to be the most widely covered aspect in the existing research with only 2 out of 12 (17%) new criteria, whereas half of the 12 criteria (50%) at the context level were not found in the previous studies. The finding supports Saracevic's argument (2000) about the lack of evaluation studies at the context level. It further implies the necessity and plausibility of assessing DLs at this level as recommended by several researchers (Bishop et al. 2003; Marchionini 2000; Marchionini et al. 2003).

Table 5.3: Criteria Distribution in the Literature Review and/or the Interview

DL leverl (total number of criteria)	Interviews & literature review	Interviews only	Literature review only
Content (20)	16	4	2
Technology (13)	9	4	1
Interface (12)	10	2	2
Service (8)	6	2	1
User (12)	8	4	1
Context (12)	6	6	0
Total (77)	55	22	7

5.7 DL Notions and Constructs

5.7.1 DL as System, Process, and Extension of Organization

In terms of their emphasis, the interviewees' verbalized DL notions can be classified into three categories: DL as system, DL as process, and DL as an extension of an organization. Among the nine interviewees, only IR3 focused only on the system,

defining a DL as "any organized collection for digital information". In contrast, the notion of a DL as the extension of an organization tended to represent the mainstream, because the other eight interviewees mentioned this concept when asked to provide a definition of a DL. IA2 defined a DL as "all of the services, collections, and activities that libraries provide, either mediated in a digital fashion or existing in a digital form…" The administrator provided an example of a DL by saying, "Rutgers itself is a large digital library. It's a collaborative portal…that has been built across numerous museums, libraries, and archives." ID2 straightforwardly reported the mission of "developing Rutgers University library's equivalent of a digital library".

Unsurprisingly, all six interviewees from RUL, including three administrators and three developers, thought of the DL as an organization. However, two researchers (IR1 and IR2) also had the same vision. In fact, IR1 considered the DL a system and also an extension of an organization. From the organizational standpoint, the researcher came up with "a type of digital library that is associated with an academic library. Thus, we have a digital library as an extension of…a regular library…And these are the types of digital libraries that have strong associations in organizations, in museums." On the other hand, from the system view, the researcher used Persius Digital Library as an example. He regarded this type of DL as being "not associated with libraries, but are in a domain."

IR2's notion of a DL also was two-fold, with one of the DL as an extension of the organization and the other of the DL as process. While the organizational notion (i.e., "a digitalized library") is straightforward and common, the notion of a DL as process implies a new way of seeing a DL. In this researcher's perspective, in addition to "a digitalized library," the DL should also be defined as "a process in which a library goes

digital,…and digital components are integrated within the mainstream, within the functions of the library," based upon the findings from her cultural studies of several national digital libraries. Likely, IA2's view of a DL as process--an activity--echoes this new view. The administrator defined the DL as "all of the services, collections, and activities that libraries provide either mediated in a digital fashion or existing in digital form." Similarly, when being asked to elaborate upon her personal experiences of DL administration, IA3 said, "One interesting area may be the human resources area: training, learning, hiring, moving people from one job to another, and how exactly we can best involve everyone…"

By looking at the DL definitions provided by the interviewees, one sees more within-group divergences for the research group than inter-group differences between the researchers and other two groups. The three researchers provided the most divergent definitions. Whereas IR3 emphasized content by saying "[the] digital library is any organized collection for digital information," IR2 provided an organizational and cultural grounded definition, which is "a digitized library." By comparison, IR1's definition was the broadest. When asked to define a DL, the researcher summarized three types of DLs, namely the DL associated with an institution (e.g., Rutgers University Library), the domain DL (e.g., ACM digital library; Persius Digital Library) as well as the commercial DL (e.g., Million Books). The major divergence could be associated with the differences among the interviewees' research areas. Although all three had DL research projects, their research foci varied among DL levels, ranging from system, interface to context. In contrast to the divergence among the researchers, the DL notions from the administrators and the developers were similar, although the emphases were slightly different.

Presumably because of their common university library affiliation, the six interviewees perceived a DL as an extension of a traditional library. In other words, any things traditional libraries have should be mirrored in digital libraries, according to the six interviewees. The interview findings provide supportive evidence to Borgman's depictions (1999) that the notion of a DL is still a subject of debate, and there are differences in perceptions between the research and practice domains.

5.7.2 DL Constructs

It should be noted that in spite of the divergence in DL notions, there was no association between the DL notion standpoint and evaluation criteria for different DL levels. For instance, IR3 had the simplest DL notion, that is "any organized collection for digital information." However, his verbalized criteria for the other DL levels' evaluation were also very rich in addition to the ones for digital content evaluation.

For general knowledge about the DL and DL evaluation, all nine interviewees agreed on the division of the six DL levels and the importance of users' opinions in DL evaluation. However, digital libraries, as the most advanced and complex information retrieval systems, might have more constructs in addition to the six levels based upon the interview data. When asked to define the DL, the interviewees frequently mentioned different *activities*, and *other stakeholders.* The interviews yielded five types of activities:

- *Developing*—"We are doing all our catalogs digitally. We are increasingly doing our ordering digitally. We build our digital collections as part of our forthcoming repository." –IA2; "…the thing we are doing is that we are creating an application that we can hook conventional library holdings…convert them into digital form and then put them into digital library storage. That's the whole process what we call the workflow management system." –ID3
- *Sharing* –"We are sharing a common architecture, we can share collections across [institutions]." –IA2
- *Marketing*—"We'll start to actively market it. We'll try to get some articles in local papers. And we'll try to get fliers out to public libraries and ask them to post it on their bulletin boards. So, we haven't made real efforts yet to publicize it to the general community."—IA2
- *Organizing, accessing & using* —"But it needs a little bit more of organization and access."–IA3; "It has to talk about how they would organize information, how people would get access to it"—

IR3; "…you can organize them in certain way, so a user can use it in whatever [way] the user prefers." –ID3

- *Managing* – "Does this collection manager have services he/she needs to manage a digital library?" –ID2; "You can also ask questions: how well is it managed the whole thing."—IR1

Among these activities, some (e.g., *organizing, accessing & using*) are common to all IR systems, others (e.g., sharing, managing) are more specific for DL settings. For this reason, IA2 included activities in her DL definition. The administrator said, "I guess my definition of [a] DL would be all of the services, collections, and **activities** that libraries provide, either mediated in a digital fashion or existing by themselves in a digital form."

Similarly, compared to conventional IR systems, DL innovation is shaped by more diverse groups of people (i.e., stakeholders). When talking about the university library Web site development, IA3 mentioned the importance of having a good Web master. ID3 also highlighted human roles in DL development. The developer told the interviewer, "The format we put in is basically the decision of management for [the] digital library. It's [a] human decision. So, it's not really a software decision." Further, he explained why the Rutgers repository as a DL was successful in his view. The developer said, "The whole Rutgers repository design--the team--we have very good leadership. We have a good cataloging [team], people with a lot of cataloging experiences, metadata schema experiences. We also have people with a lot of software development experience. The whole team, I think, is a very good team."

In general, the interview data reveals that in addition to general users, DL stakeholders might also include librarians who compile content (ID1, IA3, IR2), funders who provide financial support and make DL development possible (IR2), administrators who supervise and control the process (IA1, IA2, IA3, ID1), and developers who

technically design and develop the DL (IA2, IA3, ID1, ID2, ID3, IR2).

It was found that users' voices were given much less consideration than what researchers would like to see. Almost no user involvement was found in the DL projects described by the interviewees. Actually, some interviewees had already noted the absence of user input. ID2 confessed, "We, as digital library architects, are not doing a very good job of this." According to the developer, the only user study he did was a log analysis that captured every user search and did "an extensive analysis on the search terms people were using, the number of results they got, and these thresholds," although he agreed that a real user study "should be very interesting." The common practice with users' involvement in DL innovation was to place a feedback form within a finished DL product, and "try to encourage feedback from end users." (IA2) However, as ID2 observed, rarely users completed the form. IR2 spoke from her research findings in which funders played larger roles in DL innovations than did conventional users:

> So, donors who donate funds or funders, they shape technology, or they shape collections, for example, more than technology…Users could. But they don't have equal power. When digital libraries build users in the cases that I have looked at through empirical data I have, users are very much implied, imaginary, ideal. There might be a usability component or feedback loop that developers use. But very much it is shaped by other groups, such as project managers, technology people, strict and narrow technology people…Within [a] library context, it's really very much managed through the interests of the institution, the library…I would say there are different social groups that shape technology and that impose their own views on that technology.

There should be a consideration of the new DL notions in terms of evaluation. DL activity evaluation criteria have been part of DL context evaluation (e.g., *managerial effect*), and architecture evaluation has been considered in both context and technology evaluations (e.g., *integrity with organizational/social practice and standardization*). Similarly, the important evaluation criteria perceived by other stakeholder groups should

also be taken into account for a holistic DL evaluation in addition to conventional user-centered evaluations. These criteria should be considered at the six DL levels.

In addition to the new DL constructs raised by the interviewees, some divergences in DL level foci among different stakeholder groups have been noticed through the transcript analyses. Table 5.4 compares the top three constructs/levels among the developers, the administrators, and the researchers. The numbers in the brackets are the frequencies of the digital levels the stakeholders mentioned during the interviews. Whereas both digital content and technology are in the top list for all three stakeholder groups, not surprisingly, digital technology tends to be the predominant concern of the developers, and interface appears in the top three only for the developers.

Digital context is one of the top concerns of both the administrator and researcher groups. Although there is a higher frequency of the context concept in the researcher group than in the administrator group, the frequency distribution among the three administrators is even (44% for IA1 and 28% for IA2 and IA3) while IR2 herself contributed 68% of the context frequency within the group. Considering IR2 was the only researcher from the group whose research interests were cultural analyses of DL innovation, the skewed distribution is reasonable. Yet, it can be concluded that the administrators highlight more DL contextual attributes than do the researchers.

Table 5.4: The Top Three DL Levels/Constructs from the Three Interviewee Groups

	Administrators	Developers	Researchers
1	Content (48)	Technology (28)	Content (37)
2	Technology (27)	Content (26)	Context (25)
3	Context (18)	Interface (16)	Technology (20)

5.8 Relationship between DL Construct and Evaluation Criteria

Table 5.5 compares the total verbalized frequencies for constructs and criteria at the six DL levels. It shows that the relationship originally anticipated between DL construct and evaluation criteria was not found from the interviews. Specifically, a DL level with higher verbalized frequency was not necessarily associated with increased amount of verbalized frequency of the criteria at this level. It is hard to point out which DL level is superior, and which is inferior.

Table 5.5: Total Verbalized Frequencies of Constructs and Criteria at the DL levels

DL level	Frequency	IA1	IA2	IA3	ID1	ID2	ID3	IR1	IR2	IR3	Total
Content	Constructs	23	17	7	6	7	14	8	22	13	**117**
	Criteria	48	33	11	12	66	4	31	38	28	**271**
Technology	Constructs	14	14	6	3	10	17	2	20	4	**90**
	Criteria	24	34	13	11	59	22	12	21	23	**219**
Interface	Constructs	3	6	9	4	3	7	0	5	2	**39**
	Criteria	13	33	11	4	28	10	17	40	26	**182**
Service	Constructs	2	10	4	4	7	1	0	5	0	**33**
	Criteria	9	28	12	1	22	3	19	7	16	**117**
User	Constructs	2	8	3	7	1	0	0	7	2	**30**
	Criteria	11	40	17	11	37	8	11	32	32	**199**
Construct	Constructs	9	6	5	5	6	1	5	17	2	**56**
	Criteria	34	47	36	14	58	13	15	49	22	**288**

Chapter 6 RESEARCH FINDINGS—THE CONFIRMATION STAGE (I-IMPORTANT CRITERIA, NEW CRITERIA, INTER-GROUP SIMILARITY/DIVERGENCE)

This stage is the core phrase in the dissertation research, through which the author was able to develop the anticipated holistic DL evaluation model. The model developed was based upon the data analyses of criteria importance ratings (collected through an online survey) by 434 various stakeholder survey participants (administrators, developers, librarians, researchers and general users) from 21 countries/areas. The analyses of the data collected were able to yield: (1) sets of the top perceived DL evaluation criteria consistent with the earlier interview results; (2) a statistically proven divergence among heterogeneous stakeholder groups for some of the DL evaluation criteria; (3) lists of new criteria for the six levels of DL evaluation suggested by the survey participants, and (4) the fulfillment of the core mission of the dissertation research, which is a holistic DL evaluation model encompassing diverse stakeholder groups' perspectives at all DL levels. This chapter will only summarize the first three findings. The proposed holistic DL evaluation model will be separately illustrated and explained in Chapter 7.

6.1 The Characteristics of Survey Participants

In total, 434 survey participants finished the survey, of which the data of 431 survey participants were usable. Of these 431 survey participants, 159 (37%) self-reported as librarians as their primary roles, and 158 (37%) survey participants considered themselves general users. These two stakeholder groups constituted 74% of the total survey response. Meanwhile, the DL researchers, developers and administrators numbered 53 (12%), 36 (8%), and 25 (6%) respectively (see Figure 6.1). The difference in the group sample size is associated with the potential population differences. Usually,

the numbers of DL administrators, developers, and researchers are smaller than those of librarians and general users.

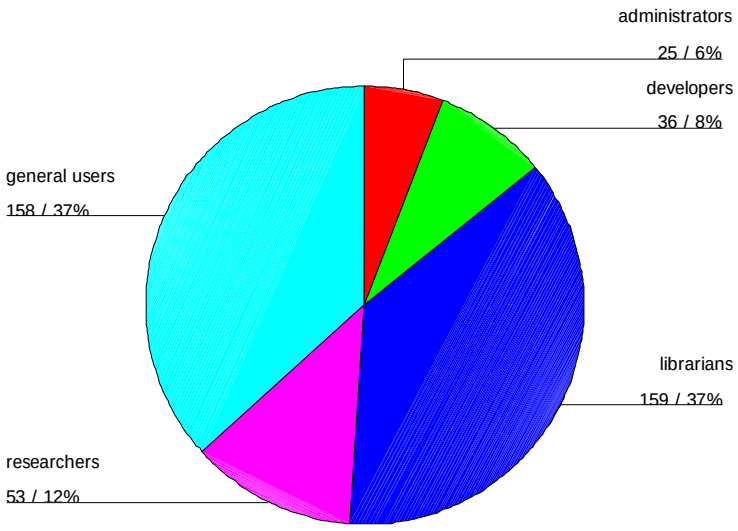

Figure 6.1: Survey Participants Distribution by DL Stakeholder Groups

All survey participants were at least 20 years old. About half of the survey participants (220, 51%) were 30 to 49 years old; 93 (22%) were over 50, and 118 (27%) were from 20 to 29 years old. Cross-tabulation between stakeholder groups and age (see Table 6.1) shows that over half (14, 56%) of the administrators were over 50, while only 14% of the developers (5) were in this age group. By comparison, the librarians had an even age distribution.

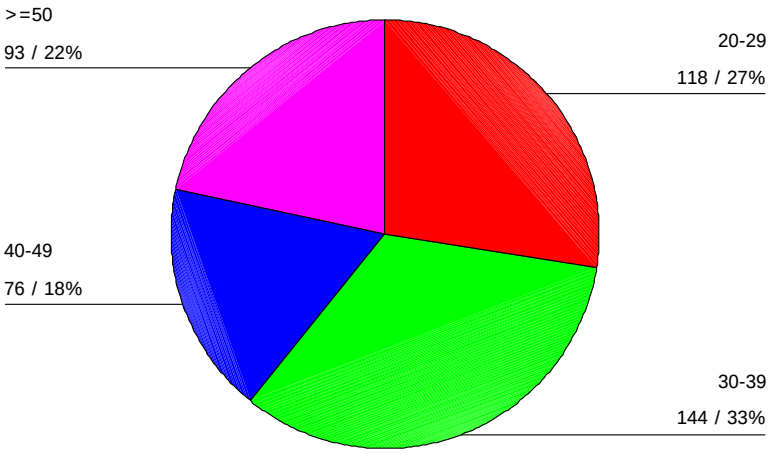

Figure 6.2: Survey Participants' Age Distribution by Years

Table 6.1: Cross-Tabulation of Survey Participants' Age x Stakeholder Group

		DL Stakeholder Groups					Total
		Administrators	Developers	Librarians	Researchers	Users	
Age in years	20-29	2	10	31	7	68	118
	30-39	4	10	59	19	52	144
	40-49	5	11	28	12	20	76
	>=50	14	5	41	15	18	93
Total		25	36	159	53	158	431

The gender distribution was 167 (38.7 %) male and 264 (61.3%) female. As Table 6.2 shows, for the administrators, developers, researchers, and general users, male and female survey participants were almost equal. However, the librarian group had many more females (114, 72%) than males (45, 28%).

Table 6.2: Cross-Tabulation of Survey Participants' Gender x Stakeholder Group

		DL Stakeholder Groups					Total
		Administrators	Developers	Librarians	Researchers	Users	
Gender	Male	11	17	45	26	68	167
	Female	14	19	114	27	90	264
Total		25	36	159	53	158	431

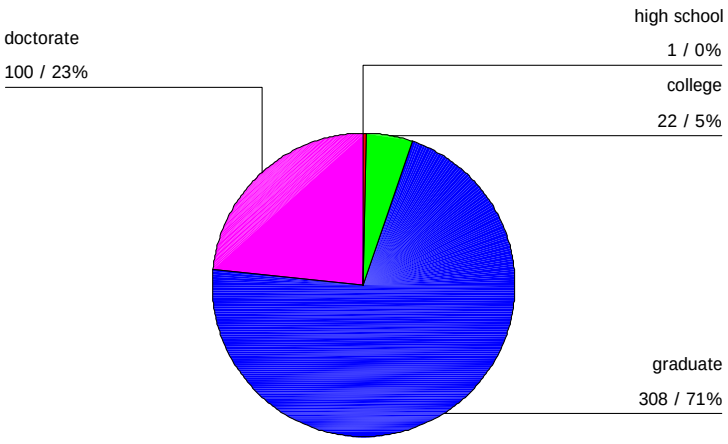

Figure 6.3: Survey Participants' Education Level Distribution

In terms of the highest education level achieved, almost all survey participants held graduate (308, 71%) or doctoral degrees (100, 23%). Only 23 survey participants (5%) had baccalaureate or lower degrees.

Table 6.3: Cross-Tabulation Survey Participants' Education x Stakeholder Group

Highest level of education	DL Stakeholder Groups					Total
	Administrators	Developers	Librarians	Researchers	Users	
High school	0	1	0	0	0	1
College	0	4	5	3	10	22
Graduate	20	24	134	21	109	308
Doctorate	5	7	20	29	39	100
Total	25	36	159	53	158	431

The subject fields show 209 (48%) for the social sciences, 130 (30%) for the sciences, 79 (185) for the humanities and arts, and 13 (3%) for others. Most survey participants (314, 73%) had been searching online for books, journal articles, and other materials for more than three years. Meanwhile, more than half of the survey participants used the Web to find resources on a daily basis (234, 54%).

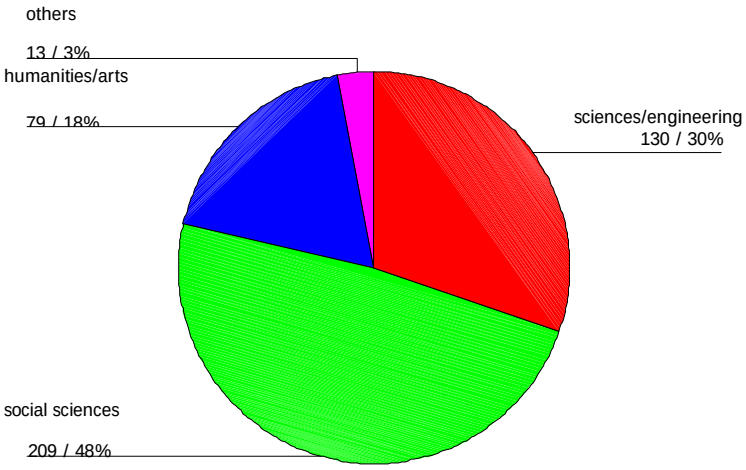

Figure 6.4: Survey Participants' Subject Field Distribution

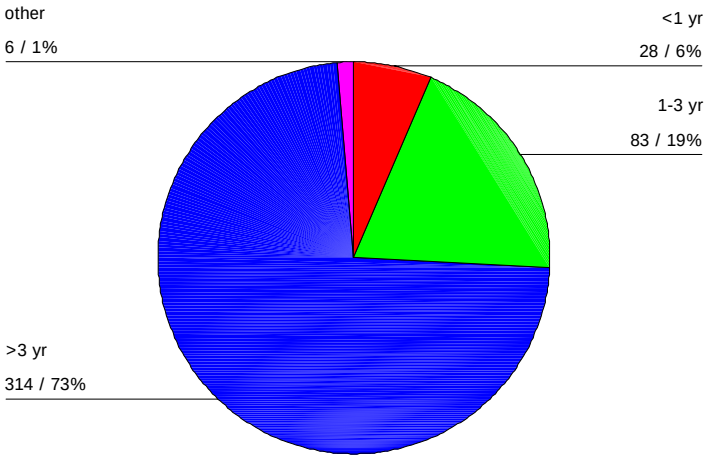

Figure 6.5: Survey Participants' Years of Online Searching Distribution

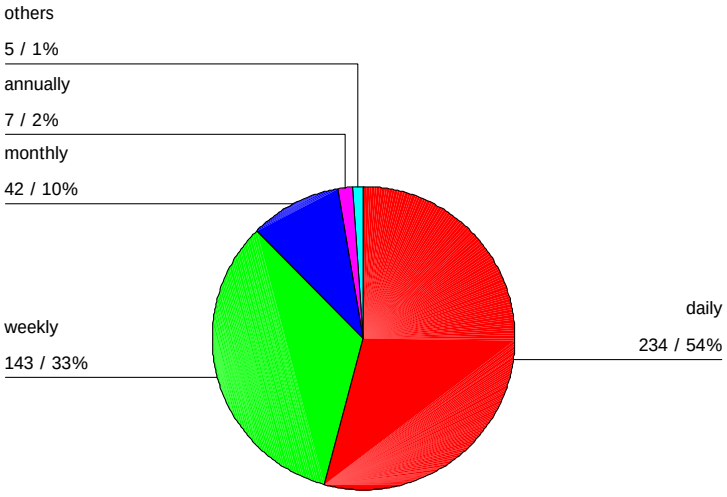

Figure 6.6: Survey Participants' Frequency of Online Searching

Table 6.4: Survey Participants' Country Origin Distribution

Country (district) origin	Frequency	Percentage
(No report of the county origin)	64	14.8
Antigua and Barbuda (West Indies)	1	.2
Australia	4	.9
Bangladesh	1	.2
Brazil	1	.2
Canada	5	1.2
China	15	3.5
Egypt	1	.2
Finland	1	.2
Germany	3	.7
Greece	3	.7
Hong Kong, SAR	1	.2
India	2	.5
Italy	1	.2
Japan	1	.2
Kenya	1	.2
Korea	1	.2
Mexico	1	.2
New Zealand	3	.7
Spain	3	.7
Sweden	1	.2
United Kingdom	7	1.6
United States	310	71.9
Total	431	100.0

Since some of the mailing lists (e.g., IFLA_L, Web4Lib), to which the author sent the call for the survey participation had international subscribers, the survey received some overseas survey participants (see Table 6.4). Among 367 (85%) survey participants who reported their nations and cities, 310 (85%) were from the United States, and only a small portion (57, 15%) were from 21 other countries as listed below. Additionally, among the 310 United States survey participants, 161 (52%) were from New Jersey, while the others are from other states, including NY (38), CA (16), PA (14), CO (9), FL (7), IL (7), GA (6), MD (5), MI (5), NC (5), TX (5), OH (3), VA (3), WI (3).

6.2 The "Don't Know" Answers

The survey participants' "don't know" answers were excluded from the statistical analysis, because they had no impact on importance. Rather, the answers could perhaps mean that the survey participants either didn't decide whether a given criterion was important or not, or they just simply had no knowledge about the criterion. Among the 51 criteria, six criteria (12%) did not receive any "don't know" answers. They are *accessibility to content,* interface *attractiveness* and *ease of use*, *technologically ease of use*, users' *satisfaction* and *successfulness of task completion.* The other 45 criteria had 1 to 27 "don't know" answers. The two criteria having more than 20 counts (5%) of the "don't know" answers are *integrity of information* (26, 6%) and *service gaps between expectation and perception* (27, 6%).

In addition, the number of "don't know" answers varied among the six DL levels. By looking at the average number of the "don't know" answers per criterion at a given DL level, the context level received the highest number of "don't know" answers (11 counts per criterion), while the interface level had the smallest number (4 counts per

criterion). This is because the survey participants were more familiar with the interface through direct interactions with DLs, but they were less concerned about the context level, which was not as closely associated with their specific information seeking experiences.

Further, the majority of the "don't know" answers were found in the general user and the librarian groups. Table 6.5 below shows the "don't know" answer distribution among the stakeholder groups. As the table shows, the general users and the librarians had 268 (72%) out of 373 total "don't know" counts. Although the actual valid cases may vary among the criteria due to the removal of various counts of the "don't know" observations, the validity of the data analyses should not be affected. Instead, the exclusion could make the results of the analysis more valid.

Table 6.5: The "Don't Know" Answer Distribution among the Stakeholder Groups

DL level (number of criteria associated)	DL Stakeholder Groups					Total
	Administrators	Developers	Librarians	Researchers	Users	
Content (9)	6	4	17	13	19	59
Technology (9)	2	7	13	3	33	58
Interface (8)	1	1	9	6	13	30
Service (7)	2	1	26	11	22	62
User (9)	10	2	20	12	18	62
Context (9)	7	4	36	13	42	102
Total	28	19	121	58	147	373

6.3 The Most and Least Important Criteria at the DL Levels

From the individual distribution pattern, the importance ratings for all criteria are negatively skewed from -.610 (IF-*Personalizability*) to –3.330 (CT-*Accuracy*). However, "In a large sample, a variable with significant skewness often does not deviate enough from normality to make a realistic difference in the analysis." (Tabachnick & Fidell 2001 p.74)

The following six tables (Table 6.6 to Table 6.11) summarize the five most important criteria as well as the lowest regarded criterion (in italics) of the six DL levels perceived by the survey participants. Meanwhile, comparisons are made between the criteria with those highly regarded earlier by the interviewees. The rankings of the importance rating are based upon descriptive data (i.e., the mean scores and standard deviation). The larger is the mean score and the smaller the standard deviation, the higher the ranking.

6.3.1 Content Level Evaluation Criteria

The top five criteria for DL content evaluation were *information accuracy, accessibility to digital content, usefulness to target users, fidelity* as to original copy, and *integrity of information*. The results are essentially consistent with the interview results. In particular, *usefulness to target users* was consistently top ranked. It appeared in the top five lists of the survey, the interview card sorting (CS), and the interview open question answering (QA). Unanimously, the interviewees and the survey participants regarded *conciseness of information* as the least important criterion for digital content evaluation. Nevertheless, within the top five lists, there was a slight variance between the two studies. For instance, *ease of understanding dropped to sixth place in the survey, while within the top fifth in CS and QA. Similarly, usefulness to target users* was perceived as the most important criterion in both QA and CS, but it dropped to third place in the survey. The slight difference is due to the inclusion of the user group in the survey, while in the interviews, there were merely administrators, developers, and researchers whose primary roles in digital libraries differ from those of general users.

Table 6.6: Content: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Accuracy of information	6.53 (1.07)		x
Accessibility	6.52 (1.00)	x	
Usefulness to target users	6.09 (1.19)	x	x
Fidelity	6.04 (1.21)		x
Integrity of information	5.97(1.17)	x	
Conciseness of information	*5.14 (1.38)*	*x*	*x*

6.3.2 Technology Level Evaluation Criteria

Digital technology evaluation also was ranked consistently between the survey and the interview results. Technology *reliability, effectiveness* and *interoperability* among systems unanimously appeared in the top lists of the survey as well as the interview CS and QA. Both *ease of use* and *efficiency* were highly rated in the survey and the interview question-answering section. However, *flexibility* was highly ranked in the interview CS but in the survey was the lowest-ranked criterion. Correspondingly, the two lowest-ranked criteria in the interview (i.e., d*isplay quality* and *security)* were ranked more highly. Likely, the underlying reason for the content evaluation (i.e., the inclusion of the user group) is also applicable here, because flexibility seems to be of greater interest to DL developers than to users.

Table 6.7: Technology: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Reliability	6.49 (0.93)	x	x
Ease of use	6.35 (1.02)	x	
Effectiveness	6.21 (1.00)	x	x
Interoperability	6.05 (1.23)	x	x
Efficiency	6.03 (1.07)	x	
Flexibility	*5.64 (1.45)*		

6.3.3 Interface Level Evaluation Criteria

All highly ranked criteria in both CS and QA--*ease of use, effectiveness,* and *appropriateness to target users*--were also ranked at the top in the survey. Users come to interact with a given interface because they want to find information needed through the interface. Therefore, it is not surprising that their top priority was the ability of the interface to support the users to find information. In order to do so, the interface has to accommodate the backgrounds, needs and behaviors of prospective users. It also has to be intuitive and transparent to users. It should be noted that *attractiveness* was still the second lowest ranked criterion in the survey despite its absence from the table.

Table 6.8: Interface: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Effectiveness	6.35 (0.99)	x	x
Ease of use	6.33 (1.02)	x	x
Consistency	5.88 (1.16)		x
Effort needed	5.88 (1.19)		x
Appropriateness to target users	5.83 (1.15)	x	x
Personalizability	*4.75 (1.46)*		*x*

6.3.4 Service Level Evaluation Criteria

The results of the survey and the interviews are consistent. In particular, *service accessibility* and *integrity to the information seeking path* appeared in the top five list of the interview CS and QA as well as in the survey. Similarly, *courtesy* was the lowest-ranked criterion in the three lists, presumably because it does not directly influence whether or not a user gets what he/she seeks. The only disagreement was *gaps between expectation and perception,* which was excluded from the top five list of the survey while appearing in both CS and QA interview top results. Again, this might be because general users usually do not pay close attention to the gap. Instead, they anticipate that a given

service can be used at the point of need without requiring them to step out from where they are; it can be trusted to provide them with fewer or no mistakes/errors, and the outcome is useful to them.

Table 6.9: Service: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Reliability	6.39 (1.00)		x
Accessibility to service	6.29 (1.09)	x	x
Usefulness to target users	6.28 (1.06)	x	
Responsiveness	6.17 (1.08)	x	x
Integrity to information seeking path	5.93 (1.17)	x	
Courtesy	*5.28(1.39)*	*x*	*x*

6.3.5 User Level Evaluation Criteria

Compared to the other DL levels, user evaluation shows a large inconsistency between the interview and the survey results. Although *successfulness* and *efficiency of task completion* as well as *productivity of users* appeared in the top lists of both the survey and the interview, *satisfaction* rose to the top of the survey list despite its ranking in the interview as one of the least important criteria. Meanwhile, some criteria that were highly regarded in the interview (e.g., learning effect and information literacy) are not in the top of the survey list. Even, *behavior changes* dropped to the lowest-ranked criterion for evaluating a DL at the user level. Again, this is presumably associated with the inclusion of user groups in the survey. Users tend to care more about the direct effects of using a DL, such as *efficiency* and *successfulness of task completion*. By comparison, *behavior changes, information literacy,* and *learning effects* are too difficult to measure. Moreover, they are likely indirect outcomes as opposed to finding information for given needs.

Table 6.10: User: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Successfulness of task completion	6.38 (0.98)	x	x
Satisfaction	6.07 (1.19)		
Efficiency of task completion	6.06 (1.07)		x
Use/reuse	6.02 (1.13)	x	
Productivity	5.94 (1.27)	x	x
Behavior change	*5.13 (1.38)*		

6.3.6 Context Level Evaluation Criteria

Similar to DL evaluation at the user level, context evaluation was less agreed-upon between the survey participants and the interviewees, in particular in the aspects of *network effect, integrity to social practice* and *extended social impact.* The interviewees and the survey participants tended to agree on *sustainability* as the most important criterion for assessing a DL at its context level. However, they were unlikely to agree on the importance of DL's *extended social impact,* including supporting multi-disciplinary research, improving the social and economic status of prospective users, preserving knowledge and culture, etc. Whereas *this criterion* was highly regarded in the interview QA, it became the least important criterion in the survey. Meanwhile, another highly ranked criterion (i.e., *integrity to social practice*) in the interview also dropped to the least important level (the least second). In contrast, incoming and outgoing hyperlinks (i.e., *network effect*) in the survey participants' perspectives were likely to be important to a certain extent, while it was the lowest-ranked criterion in the interview QA and CS.

Table 6.11: Context: Survey Participants' Top Five and Least Important Criteria (in Italics)

Criteria	Top five and the lowest-ranked criteria in the survey Mean (SD) n=431	Inclusion in the interview top five and the least lists	
		QA	CS
Affordability/sustainability	6.32 (1.05)	x	x
Collaboration/sharing	5.92 (1.10)	x	
Copyright compliance	5.82 (1.58)		x
Managerial support	5.76 (1.23)	x	
Network effect	5.66 (1.26)		
Extended social impact	*5.19 (1.41)*		

6.3.7 The Combined Most and Least Important Criteria Across the Six DL Levels

After merging these six tables into a single table and ranking the 51 criteria by their means and then standard deviations, the author drew up a combined list of the most important criteria (Table 6.12) and a combined list of the least important criteria (Table 6.13). The top ten criteria are composed of more criteria for the lower DL level evaluation (i.e., content, technology, interface, service) as opposed to the higher-level evaluation: user and context. Specifically, the top list includes two service criteria, two content criteria, two technology criteria, two interface criteria, one user criterion and one context criterion.

Table 6.12: The Combined Top Ten Criteria for the Six DL Levels (n=431)

Criteria	Mean	SD
CT-Accuracy of information	6.53	1.07
CT-Accessibility	6.52	1.00
TN-Reliability	6.49	0.93
SV-Reliability	6.39	1.00
UR-Successfulness of task completion	6.38	0.98
IF-Effectiveness	6.35	0.99
TN-Ease of use	6.35	1.02
IF-Ease of use	6.33	1.02
CX-Affordability/sustainability	6.32	1.05
SV-Access to service	6.29	1.09

In contrast, the list of least-important criteria contains more higher-level evaluation criteria, including four context criteria, and one user criterion along with two

service criteria, two interface criteria, one content criterion and one user criteria. Clearly, the context evaluation criteria on average received the lowest importance ratings in comparison to those for the other five digital aspects. Only one context criterion (*sustainability*) appears within the top ten mixed criteria, whereas four of them (i.e., e*xtended social impact, integrity to organizational/social practice, productivity of community members,* and *outcome against organizational goals*) dropped into the least ten mixed criteria. This implies that to date the significance of a DL's impacts on institutional, social, cultural, and other contexts have not been well studied. Or, it just simply means that DL users and other stakeholders are more concerned about the DL aspects with which they are directly dealing. For instance, the content and service aspects are the primary concerns in DL innovation. Two (i.e., *reliability and accessibility*) out of the seven service evaluation criteria (29%) are in the top ten list. Likewise, among the nine content evaluation criteria, *accessibility* and *accuracy of information* are the top first and second most important criteria, respectively, in the mixed list.

Table 6.13: The Combined Least Ten Criteria for the Six DL Levels (n=431)

Criteria	Mean	SD
IF-Personalizability	4.75	1.46
UR-Behavior change	5.13	1.38
CT-Conciseness of information	5.14	1.38
CX-Extended social impact	5.19	1.41
SV-Courtesy	5.28	1.39
IF-Attractiveness	5.29	1.30
CX-Integrity to social practice	5.42	1.33
CX-Productivity of community members	5.47	1.42
SV-Gaps between expectation and perception	5.53	1.22
CX-Outcome against organizational goals	5.55	1.39

In general, these DL stakeholders are concerned about the ability to access high quality content and service (**the Premise)**. Their second concern is ease of search and use

during their interaction with the content and service (**the Process**). Then, they value **the direct Performance** of using the DL, such as being able to find more relevant resources (e.g., *effectiveness*), and *successfulness of task completion*. In contrast, the least perceived criteria are the indirect outcomes of DL use (i.e., not directly related to finding expected information, such as. behavior change), or non-core processes and premises, such as *personalizability of an interface, courtesy of a service, conciseness of information*, etc.. However, in general the least perceived criteria have a larger standard deviation within the sample than the most perceived criteria, which might be an implication that these criteria received divergent important rankings among the stakeholders.

6.4 Similarity/Divergence among DL Stakeholder Groups

Not all 51 DL evaluation criteria have statistically significant differences among the five DL stakeholder groups. ANOVA results show that only 11 out of 51 criteria (less than 22%) have statistically significant differences among the stakeholder groups regarding how important these criteria are in assessing a DL. Table 6.14 summarizes the ANOVA results. Service, interface and user evaluation criteria received much more consensus among the groups on the importance ratings, which is in line with the interview results. In contrast, the context evaluation criteria are the ones that have the most group divergence. This might be another reason for the lower importance rankings for some of the context evaluation criteria.

Further, Scheffe's post-hoc test (good for unbalanced sample sizes) results show that the differences exist only among some (not all) of the five stakeholder groups. For instance, the group difference on the perceived importance of *content appropriateness to target users* was only found between the administrators and the general users. In addition,

the differences existed primarily between general users and the other stakeholder groups, including the administrators (6 criteria), the librarians (8 criteria), and the researchers (2 criteria). What the administrators, librarians and/or researchers sometimes highly perceived were usually the ones that were least regarded by the users. For instance, unlike the other stakeholder groups' perspectives, all *appropriateness* criteria for the aspects of digital content, technology, and interface were not favored by the general users. Whereas the administrators and the librarians regarded *copyright compliance* and other context level evaluation criteria, the general users tended to hold the opposite view. However, *comprehensiveness of collection* was the only criterion that had much higher rankings from general users. Interestingly, no significant effect was found between the developer group and any of the other four groups.

Table 6.14: DL Evaluation Criteria with Statistical Inter-group Divergence (n=431)

Criteria	ANOVA results	Groups with sig. difference (mean difference, α)
CT-Appropriateness to target users	F(4,423)=3.889, p<.005	Administrator -user (.78, .05)
CT-Comprehensiveness of collection	F(4, 425)=5.048, p<.001	Librarian – user (-.53, .005)
TN-Appropriateness to digital information	F(4,410)=4.136, p<.005	administrator -user (.80, .05); Librarian – user (.46, .05)
TN-Interoperability	F(4,415)=4.042, p<.005	Librarian – user (.47, .05)
TN-Security	F(4,423)=3.618, p<.01	administrator-user (.84, .05)
IF-Appropriateness to target users	F(4,424)=8.116, p<.001	administrator-user (.95, .005); librarian–user (.54, .001); researcher-user (.72, .005)
UR- Acceptance	F(4,421)=3.991, p<.005	librarian–user (.42, 05)
CX-Copyright compliance	F(4,416)=6.753, p<.001	administrator-user (1.09, .05); librarian–user (.82, .001)
CX-Extended social impact	F(4,410)=3.646, p<.005	researcher-user (.71, .05)
CX-Integrity to organizational practice	F(4,414)=4.057, p<.005	librarian–user (.51, .05)
CX-Managerial support	F(4,416)=5.152, p<.001	administrator-user (1.00, .05); librarian–user (.45, .05)

In addition to the statistically significant effects for individual evaluation criteria, the group difference can also be found through comparing the top ranking criteria among

the stakeholder groups. The Table 6.15 to Table 6.20 below compare the top five highly perceived criteria among the five stakeholder groups at each of the six DL levels. From these tables, for a given DL level evaluation, some criteria are on the top five lists from all stakeholder groups (see text in bold), while the others are perceived as being important by some of the groups. For instance, content evaluation has three criteria (i.e., *accessibility, accuracy,* and *usefulness)* that were perceived to be important by all five groups of stakeholders. However, the administrators considered *appropriateness* and *integrity of information* more important than *ease of understanding*, which was on the top five lists of the other four stakeholder groups, but not on the administrators' list. Nevertheless, none of the four groups rated *integrity of information* as important as did the administrator group. Additionally, *comprehensiveness* and *fidelity of information* only show up in the users' and developers' top five lists respectively.

Table 6.15: Comparison of the Top Five Criteria among the Five Stakeholder Groups (Content)

Criteria	Administrators (n=25)	Developers (n=36)	Librarians (n=160)	Researchers (n=53)	Users (n=157)
Accessibility	**X**	**X**	**X**	**X**	**X**
Accuracy of information	**X**	**X**	**X**	**X**	**X**
Usefulness to target users	**X**	**X**	**X**	**X**	**X**
Ease of understanding		X	X	X	X
*Appropriateness to target users	X		X	X	
*Comprehensiveness					X
Fidelity		X			
Integrity of information	X				

* criteria statistically proven to have an inter-group difference.

Similarly, for technology evaluation, all stakeholder groups regarded *ease of use* and *reliability* as two of the most important criteria. However, unlike the administrators, developers, and librarians, the researchers and general users tended to overlook the

system's capability to protect the system as well as a user's personal information (i.e., the *security* issue). Instead, the latter two groups of stakeholders along with the developers were more concerned with the issue of *efficiency* of digital technology. Only the general users noted the importance of *display quality.*

Table 6.16: Comparison of the Top Five Criteria among the Five Stakeholder Groups (Technology)

Criteria	Administrators (n=25)	Developers (n=36)	Librarians (n=160)	Researchers (n=53)	Users (n=157)
Ease of use	**X**	**X**	**X**	**X**	**X**
Reliability	**X**	**X**	**X**	**X**	**X**
*Interoperability	X	X	X	X	
Effectiveness	X		X	X	X
*Security	X	X	X		
Efficiency		X		X	X
Display quality					X

* criteria statistically proven to have an inter-group difference.

The general users tended to opt for an interface that required less *effort*. They did not put *appropriateness to target users* on their top list. Meanwhile, only the researcher group thought *supportiveness of HCI* was unimportant. Nevertheless, all five groups agreed on the importance of *ease of use, effectiveness,* and *consistency* for DL interface evaluation.

Table 6.17: Comparison of the Top Five Criteria among the Five Stakeholder Groups (Interface)

Criteria	Administrators (n=25)	Developers (n=36)	Librarians (n=160)	Researchers (n=53)	Users (n=157)
Ease of use	**X**	**X**	**X**	**X**	**X**
Effectiveness	**X**	**X**	**X**	**X**	**X**
Consistency	**X**	**X**	**X**	**X**	**X**
*Appropriateness	X	X	X	X	
Interaction support	X	X	X		X
Effort needed				X	X

* criteria statistically proven to have an inter-group difference.

There was no divergence in the five stakeholders' perspectives regarding the top five digital service evaluation criteria.

Table 6.18: Comparison of the Top Five Criteria among the Five Stakeholder Groups (Service)

Criteria	Administrators (n=25)	Developers (n=36)	Librarians (n=160)	Researchers (n=53)	Users (n=157)
Accessibility to service	**X**	**X**	**X**	**X**	**X**
Integrity to information seeking path	**X**	**X**	**X**	**X**	**X**
Reliability	**X**	**X**	**X**	**X**	**X**
Responsiveness	**X**	**X**	**X**	**X**	**X**
Usefulness to target users	**X**	**X**	**X**	**X**	**X**

While the general users and researchers were more concerned about their *productivity* in research, work, and daily life, the DL administrators, developers and librarians cared more about direct accountability of a DL, which is *use/reuse*. Compared to these two criteria with disagreement among the groups, *effectiveness (i.e., successfulness)* and *efficiency of task completion* and satisfaction were unanimously regarded as the important criteria.

Table 6.19: Comparison of the Top Five Criteria among the Five Stakeholder Groups (User)

Criteria	Administrator (n=25)	Developer (n=36)	Librarian (n=160)	Researcher (n=53)	User (n=157)
Successfulness of task completion	**X**	**X**	**X**	**X**	**X**
Satisfaction	**X**	**X**	**X**	**X**	**X**
Efficiency of task completion	**X**	**X**	**X**	**X**	**X**
Use/reuse	X	X	X		X
*Acceptance	X	X	X	X	
Productivity				X	X

***** criteria statistically proven to have an inter-group difference.

Although *managerial support,* "the extent to which digital library development and maintenance is supported by human and/or physical resources, or vice versa," had statistically proven group difference, it was perceived as one of the important DL context evaluation criteria by all five stakeholder groups. Additionally, *sustainability of digital libraries* and *collaboration/sharing among DLs* were the other two criteria that had received inter-group consensus. However, *copyright compliance* was perceived

differently by the general user group as compared to the other four groups. The users were unlikely to care about this issue. Rather, they thought *productivity of community members* was more important. Interestingly, only the researcher group saw the importance of DL influences on higher social, economic, political, cultural aspects.

Table 6.20: Comparison of the Top Five Criteria among the Five Stakeholder Groups (Context)

Criteria	Administrators (n=25)	Developers (n=36)	Librarians (n=160)	Researchers (n=53)	Users (n=157)
Affordability/sustainability	**X**	**X**	**X**	**X**	**X**
Collaboration/sharing	**X**	**X**	**X**	**X**	**X**
***Managerial support**	**X**	**X**	**X**	**X**	**X**
*Copyright compliance	X	X	X	X	
Network effect	X				X
Outcome against organizational goals		X	X		
*Extended social impact				X	
Productivity of community members					X

* criteria statistically proven to have an inter-group difference.

Clearly, the service evaluation had the largest consensus (100% agreement) among the stakeholder groups, and the technology evaluation received the most divergent opinions (29% agreement) with respect to the five most highly ranked criteria. The agreement rates for the other four DL level evaluations were 37% for the content and the context, and 50% for the interface and the user. It should be noted that the lower agreement for the technology evaluation was also found in the interviews. The underlying reason is associated with the unfamiliarity of DL technology by the majority of the stakeholders except the developers.

It should be noted that DL stakeholder groups' impact on the DL criteria ratings might also be affected by other demographic factors, such as age, gender, education level, major, and online searching experience. For instance, univariate analysis of variance reveals a significant interaction ($F_{(18, 382)}$=2.442, α=.001) between the survey participants'

stakeholder group affiliation and online searching frequency on their perceived importance of content *appropriateness* to target users. Since addressing the other factors' effects are beyond the dissertation research objectives, and the size of the research, the interaction effects are purposefully omitted from the dissertation and will be discussed later in a separate study.

6.5 New Criteria Revealed by the Survey Participants

There are a few "new" criteria identified from the open-end survey questionnaire session. Some of these criteria appeared in the earlier exploratory stage (the interview) findings and were not included in the survey questionnaire due to their lower rankings/frequencies. For instance, one developer suggested that "whether information is not already available online"(*uniqueness of resource*) could be used to judge the quality of digital content. Other survey participants also mentioned *diversity of information* ("multiple formats of content"), *cost* and *accessibility* of required software/hardware, interface *clarity*, usability related *error-handling* capability, and users' *feedback* on service received.

Meanwhile, there were a few criteria that are truly new with respect to the two research stages. One user suggested that (collection, feature, service) *awareness to users* should be considered. Echoing the user's perspective, one administrator noted the "importance of communicating to users what assistance services are offered." Similarly, one researcher proposed that it is necessary to do some *marketing* for a given DL. In addition to the *size of a collection*, a librarian regarded *the size of files* within the collection was also "somehow significant." Meanwhile, users also wanted additional

service availability, such as the "ability for users to create and annotate content to enhance the DL's value."

Interestingly, unlike earlier interview findings in which more new criteria were identified for higher level (e.g., context) DL evaluation, the survey results show more new criteria for lower level (e.g., content, technology, and interface) evaluations.

Although these new criteria are promising and worth further investigation, each of them only represents personal perspectives from a few out of the 431 survey participants, and thus would not be included in the proposed evaluation model (see Chapter 7) and the experiment instruments in the verification stage.

Chapter 7 RESEARCH FINDINGS—THE CONFIRMATION STAGE (II-THE PROPOSED HOLISTIC EVALUATION MODEL)

This chapter reports the core findings of the dissertation research, which is the proposed holistic DL evaluation model. Although the model was directly constructed by analyzing the 431 cases of the online survey data, the actual model development should be traced back to the earlier dissertation research stage (i.e., exploration) in which various existing and potential evaluation criteria were identified and examined through a literature review and an in-depth interview with nine DL innovation experts. The 77 criteria identified from the exploration stage indirectly served as the base for the construction of the final model. These criteria were used to develop an online survey questionnaire in the confirmation stage. After applying descriptive and inference statistical techniques to the analysis of 431 sets of finished survey data from five groups of DL stakeholders (i.e., administrators, developers, librarians, researchers, users), the author was able to obtain 37 highly perceived criteria and use them to construct the evaluation model.

The model contains 19 core and 18 case-by-case criteria. It is based on certain rules to determine what criteria should be included in the model, and to which core or case-by-case category a given criterion should be assigned. The criteria with high importance rankings from the five groups of stakeholder groups as well as consensus among them were added to the holistic model for DL evaluation as core evaluation criteria. Meanwhile, those key criteria with lower agreement rates were selectively included as case-by-case criteria based upon certain pre-defined rules. First, the case-by-case criteria should be those that have statistically proven differences among the groups

regarding their value as indicators in DL evaluation. Then, for those with no significant effects according to the post-hoc results, they should meet either of two conditions before being included in the model: (1) be within the top three of a given stakeholder group (see Table 6.15 to Table 6.20 in the 6.2.4 section); and/or (2) be on the top five list of a given DL level (see Table 6.6 to Table 6.11 in the 6.2.3 section). Under these rules, for instance, *display quality* was on the general users' top five list (the fifth ranking), but it was excluded because it is not on the combined top five lists for technology evaluation. In general, the "core" notion suggests that each DL should be evaluated with these criteria, and the "case-by-case" notion implies stakeholder interest-based selectivity.

The following two sections present two versions of the proposed DL evaluation model, namely tabular and graphic. Following the two sections will be the further elaboration section on the proposed model with focus on (1) what criteria are included as core as opposed to case-by-case, and (2) what implications these criteria hold for DL evaluation.

7.1 The Tabular Presentation of the Model

Table 7.1 is the tabular presentation of the proposed holistic DL evaluation model. It displays the six sets of evaluation criteria corresponding to the six DL levels as described in Saracevic's stratified IR model (1996, 1997): content, technology, interface, service, user, and context. For each DL level, there are two groups of criteria with one group for the core (i.e., the criteria with high importance ranking consensus among the five stakeholder groups) and the other for the case-by-case (i.e., the criteria receiving high importance rankings from some but not all stakeholder groups). The second column lists the 19 core criteria. Each cell in the column maps to a given DL level. For example,

the core evaluation criteria for digital content evaluation include *accessibility to collection/information*, *information accuracy*, and *usefulness to target users*. The core criteria for digital technology evaluation are *reliability* and *ease of use*.

Table 7.1: The Proposed Holistic DL Evaluation Model (Tabular Presentation)

DL Level	Core Criteria	Case-by-Case Criteria by Stakeholder Groups*					
			USR	RES	LIB	DEV	ADM
Content	Accessibility Accuracy Usefulness	Appropriateness		X	X		X
		Comprehensiveness	X				
		Ease of understanding	X	X	X	X	
		Fidelity				X	
		Integrity					X
Technology	Ease of use Reliability	Effectiveness	X	X	X		X
		Efficiency	X	X		X	
		Interoperability		X	X	X	X
		Security			X	X	X
Interface	Ease of use Effectiveness Consistency	Appropriateness		X	X	X	X
		Effort needed	X	X			
		Interaction support	X		X	X	X
User	Successfulness Satisfaction Efficiency	Acceptance		X	X	X	X
		Productivity	X	X			
		Use/reuse	X		X	X	X
Context	Sustainability Collaboration Managerial support	Copyright compliance		X	X	X	X
		Extended social impact		X			
		Network effect	X				X
Service	Accessibility Integrity Reliability Responsiveness Usefulness	No case-by-case criteria					

*USR-user, RES-researcher, LIB-librarian, DEV-developer, ADM-administrator

The third column shows the 18 case-by-case criteria, which can be mapped to various DL levels. In the bottom row, there are no case-by-case criteria for the evaluation of digital service; only five core criteria are included, this is because these criteria have been unanimously agreed upon by all five stakeholders as important. Section 7.3. has a detailed discussion of the distribution pattern and its implication for DL evaluations. In combination with the third column are the case-by-case criteria distributions among the

five stakeholder groups (the five right columns): general users (USR), researchers (RES), librarians (LIB), developers (DEV), and administrators (ADM). The "X" indicates the inclusion of a given criterion on a given stakeholder group's top ranking list. For example, according to the model, for interface evaluation, general users tend to ignore *appropriateness* to their background, but highlight *effort needed* and *supportiveness of HCI* . The three librarianship domain groups (i.e., administrators, developers and librarians) have the opposite opinion, which favors the appropriateness and devaluing the effort needed. A detailed illustration showing how the case-by-case criteria are distributed among the groups can be found in section 7.3.

7.2 The Graphic Presentation of the Model

The graphic model is derived from the tabular presentation of the model. While the tabular presentation is intuitive and easy to understand, the graphic version of the model (see Figure 7.1 below) is more meaningful by clearly differentiating the core and non-core criteria through distinct areas (center versus outer rings) in concentric circles as well as the relationship of the six levels to DL innovation. The combination of the two model representations strives to improve the understanding of the model. Appendix 5 lists the definitions for these criteria.

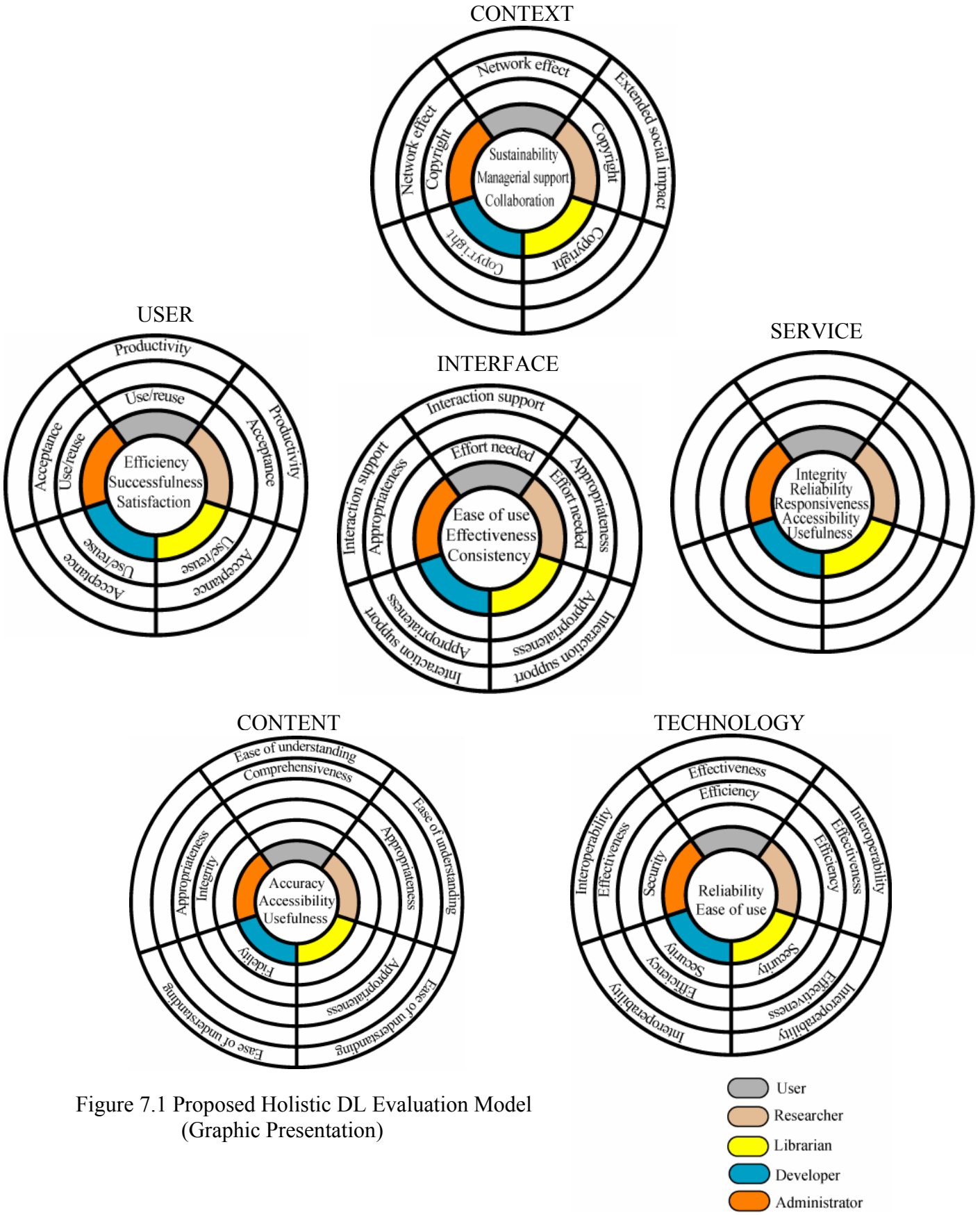

Figure 7.1 Proposed Holistic DL Evaluation Model (Graphic Presentation)

The graphic version of the model is composed of six sets of concentric circles. Each set contains important criteria at a given DL level, from the context significant criteria at the top reflecting the context as the highest DL level as suggested by Saracevic (2000); the content and technology at the bottom representing the two fundamental components in a DL, and the interface in the middle demonstrating its central position in a DL where the other DL level components meet. The user and service circles, representing the two DL levels with human users' and agents' involvement, are left and right of the interface circle respectively. The text above a circle indicate which DL level it represents. Within a single concentric circle, where the criteria in the center are core criteria with consensus from all the five stakeholder groups, the ones in the radiated outer rings are specially selected for meeting various groups' interests for tailored DL evaluation purposes (case-by-case criteria marked with five different colors for the five groups). The color key at the right bottom shows the colors of the stakeholder groups, such as silver for general users, light brown for researchers, yellow for librarians, blue for developers, and orange for administrators. Each outer ring contains a single criterion that has been perceived to be important by at least one group but less than five groups of stakeholders.

The number of the concentric outer rings indicates the degree of inter-group divergence. The more rings, the more inter-group divergence a given DL level has regarding what should be evaluation at the level. For instance, the content circle has five outer rings with five different criteria, and the service circle has no outer rings.

7.3 Further Elaboration on the Model

The following sections describe the model, starting from the fundamental DL levels (i.e., content, technology and interface) to the higher levels (i.e., service, user, and context) that are illustrated in the graphic model. These sections will also briefly discuss the implications for DL evaluation.

7.3.1 Content Level Evaluation Criteria

The Content concentric circle (the bottom left) demonstrates the important criteria for digital content evaluations, including digital information, meta-information, and collections. For the Content evaluation, there are three core evaluation criteria and five case-by-case criteria. The model suggests that all digital content should be evaluated in terms of the extent to which they are readily *accessible* (financially, physically, technologically, legally, etc), *accurate* without noticeable errors (e.g., typos, false optical character recognition, incorrect information), and *useful to target users* in achieving certain goals. Meanwhile, the model also implies that digital content evaluation could be tailored by adopting the case-by-case criteria in the outer rings based upon the stakeholder who will benefit from the evaluation results. For instance, a user-centered digital content evaluation should include *ease of understanding* of information and *comprehensiveness* of collection as criteria. In contrast, given the evaluation report addressed to administrators, *integrity of information/collection* and *appropriateness to target users* should be highlighted. An ideal evaluation should include both core and case-by-case criteria in the model. However, there is frequently a restriction (e.g., evaluation session length) on the number of criteria included. If this is the case, the group sensitive case-by-case criteria could serve as a basis for selection.

Compared to the criteria for other than content evaluation, the evaluation criteria at the Content level seem to have larger variance among the groups. Except for the researcher and librarian groups, whose criteria (i.e., *ease of understanding* and *appropriateness to target users*) are shared with some other groups, the remaining three groups have their own unique criteria, including *comprehensiveness* of collection from the general user group, *integrity of information* / collection from the administrator group, and information *fidelity* from the developer group.

7.3.2 Technology Level Evaluation Criteria

The Technology concentric circle (the bottom right) summarizes the important criteria for digital technology evaluations, including hardware and software. In total, there are four case-by-case criteria and two core criteria for DL evaluations at the technology level. The model suggests that *reliability* and *ease of use* are the two important criteria that should be addressed in every digital technology evaluation, because both are on the top-five list of all five stakeholder groups. These two become the core criteria. Meanwhile, although in the outer rings, *interoperability* among systems and *effectiveness* might also need to be well addressed, because they both have inter-group agreement as important digital technology evaluation criteria from a large percentage of the groups (four out of five). Additionally, unlike the ones for content evaluation, all case-by-case evaluation criteria share agreement among more than three stakeholder groups.

Following the case-by-case criteria selection rules (see the first paragraph of this chapter), only two users' case-by-case criteria have been included in the model. They are *effectiveness* and *efficiency.* Effectiveness is shared with other stakeholder groups except

the librarians, whereas efficiency is on the researchers' and developers' outer rings. Likely, for the digital technology evaluation criteria, the user and researcher groups have more agreement than among the remaining groups. The only exception is *interoperability*, which is not on the users' list but on the researchers' list.

Interestingly, the three groups from the librarianship domain (i.e. administrators, developers, and librarians) tend to have more agreed-upon perspectives. In addition to the two core criteria, they all regard *interoperability* and *security* as being important. The only difference is associated with whether *effectiveness* or *efficiency* should be taken into account more seriously. Whereas the developers tend to opt for *efficiency*, the administrators and librarians are more concerned with *effectiveness* (the extent to which a digital technology has been developed to meet requirements).

7.3.3 Interface Level Evaluation Criteria

The Interface concentric circle (the middle) shows the important criteria for evaluating DL interfaces, including various features and functions the interfaces provide. In total, at the interface level, there are three case-by-case criteria and three core criteria for DL evaluations. The three core criteria for the DL interface evaluation are *ease of use, consistency,* and *effectiveness* (the extent to which a digital interface helps users find information they need). In addition to the core criteria, two case-by-case criteria, s*upportiveness to HCI* and *appropriateness to target users,* also received high inter-groups consensus. For each of the two, four out of the five stakeholder groups agreed on its importance to DL interface evaluations. General users tend not to think much about the *appropriateness* to them, while the other four groups are more sensitive to this issue. This finding is similar to the earlier findings that content *appropriateness* has also been

ignored by general users. In contrast, researchers tend to give a higher degree of importance to *appropriateness* over *interaction support*, which deviates from the remaining four groups.

Although general users and researchers have divergent opinions on the interaction support and appropriateness, both groups agree on perceiving *effort needed* as an important criterion for digital interface evaluation. Actually, they are the only two groups that include *effort needed* in their top lists. Interestingly, the model again implies that the three groups from the librarianship professional domain (administrators, developers, librarians) tend to agree the most in their perspectives. Unanimously, the three groups regarded *interaction support, appropriateness, ease of use, effectiveness,* and *consistency* as the top five DL interface evaluation criteria.

7.3.4 Service Level Evaluation Criteria

The Service concentric circle (the middle right) demonstrates the important criteria for assessing digital service, which aims to provide DL users with additional on-demand assistance, such as reference, tutorials, term suggestion, SSD-selective document dissemination, etc. The most noticeable feature in the circle is the blank outer rings, which indicates no case-by-case criteria for digital service evaluation. In fact, throughout the entire study, all stakeholder groups agreed on the five top evaluation criteria: *service accessibility, reliability, responsiveness, usefulness to target users,* as well as *integrity to information seeking path*. Accordingly, all five criteria become the core criteria. In other words, digital service evaluation should address the five issues, and the outcomes of the evaluation adopting the five criteria should be able to reflect the needs of all five groups.

7.3.5 User Level Evaluation Criteria

The User concentric circle (the middle left) summarizes the important criteria for assessing DL indirectly from its users' attributes, such as their success and efficiency of task completion, satisfaction, etc.. In total, there are three case-by-case criteria and three core criteria for DL evaluations at the user level. The circle shows that the five stakeholder groups regarded *efficiency* and *successfulness of information seeking task completion,* as well as *users' satisfaction* as three highly ranked criteria. These three became the core evaluation criteria at the user level. Meanwhile, *acceptance* and *use/reuse* were another two criteria that were widely perceived to be important among the stakeholder groups. Again, for each of the two criteria, four out of the five groups rated it as their highly important criterion. Only the researcher group did not include *use/reuse* in the top list, whereas all other four groups except the general users perceived *acceptance* as one of the five top important criteria for DL evaluation at the user level.

Again, only the two non-professional librarianship groups (i.e., general users and researchers) concur on user *productivity* as being highly important to DL success. None of the three professional librarianship groups (i.e., administrators, developers, and librarians) included this criterion in their top list. Instead, the three professional groups again shared case-by-case criteria, *acceptance* and *use/reuse.*

7.3.6 Context Level Evaluation Criteria

The Context concentric circle (the top) suggests the important criteria for assessing DL from the following two dimensions: (1) how well a given DL fits into a larger contextual (e.g., institutional, social, cultural, economic, legal) practices, and (2) what impacts and effects the DL may have on these contextual practices. In total, there

are three case-by-case criteria and three core criteria for DL evaluations at the context level. The model indicates that *sustainability, collaboration,* and *managerial support* are the three core criteria for DL evaluation at the context level. Meanwhile, *copyright compliance* is almost unanimously perceived, except by the user group, to be a very important criterion. Besides the *copyright compliance*, two other case-by-case criteria are *network effects* with other resources, and *extended social impact*.

Similar to the digital content evaluation, the Context level also has more scattered case-by-case evaluation criteria. For instance, only the researcher group perceived the extended social impact as being important, and only the administrator and general user groups highlighted the network effects. However, the difference between the two DL levels is that the majority of the stakeholder groups (i.e., general users, librarians and developers) have only one case-by-case criterion for each at the context level, while there are at least two case-by-case criteria for each group at the content level. This implies that some DL stakeholder groups tend to care less about DL evaluation at the context level in comparison to their perceptions on evaluation criteria for the remaining five DL levels. Furthermore, the lower concern primarily exists in the librarian, developer, and general user groups, considering that they have fewer case-by-case important criteria. In contrast, the administrator and researcher groups tend to pay more attention to the contextual effects of DLs.

In sum, except with the digital service evaluation whereby all important criteria are core and no case-by-case criteria have been identified, DL evaluations at the other five levels have two to three core evaluation criteria, and three to five case-by-case criteria, as the model demonstrates. In light of the model, a DL evaluation can be very

flexible, focusing on one to several DL levels and with selective case-by-case criteria, depending on evaluation objectives and the target stakeholder groups' interests in the DL evaluation results. It should be re-emphasized that a DL evaluation can be conducted by adopting the criteria in any of the six concentric circles based on evaluation objectives. However, to get a holistic picture of a DL, it is essential to assess it at various levels by examining all core and case-by-case criteria at these levels. Nevertheless, the evaluation should not necessarily fit into a single study. Instead, a final holistic picture about the DL can be drawn through an integration of the findings of several evaluations.

The holistic nature of the model is reflected in the following two aspects. First, the model incorporates diverse viewpoints from different DL stakeholder groups. Not only does the model include the core criteria perceived unanimously as important by all the five groups, but also it contains the case-by-case criteria on the top list of some stakeholder groups. The inclusion of the case-by-case criteria is nicely in line with Marchionini's (2000) multifaceted DL evaluation framework, which emphasizes different viewpoints from different stakeholders. Such an evaluation model is capable of reflecting different stakeholders' needs of all kinds. Second, the model comprises the important evaluation criteria at all six DL levels described in Saracevic's (1996, 1997) stratified model, not only reflecting the narrower sense of information retrieval systems (e.g., content, technology, and interface), but also embracing broader system components (e.g., service, user, and context). Hence, the proposed model might be able to discover how well a given DL is developed, while still being flexible in conducting DL evaluations at individual levels.

Chapter 8 RESEARCH FINDINGS—THE VERIFICATION STAGE

8.1 The Experiment Participants

There were 33 participants who participated in the experiment. Of these, 11 (33%) self-reported as general users, and 7 (21%), 6 (18%), 5 (15%), and 4 (12%) reported themselves as librarians, developers, researchers and administrators respectively. The stakeholder group identification was based upon the participants' self-reports in the Pre-Searching Questionnaire rather than the author's initial sampling frame. However, the two are almost the same, except one case where a potential DL researcher self-reported to be a DL general user. The participant was eventually categorized as a DL general user.

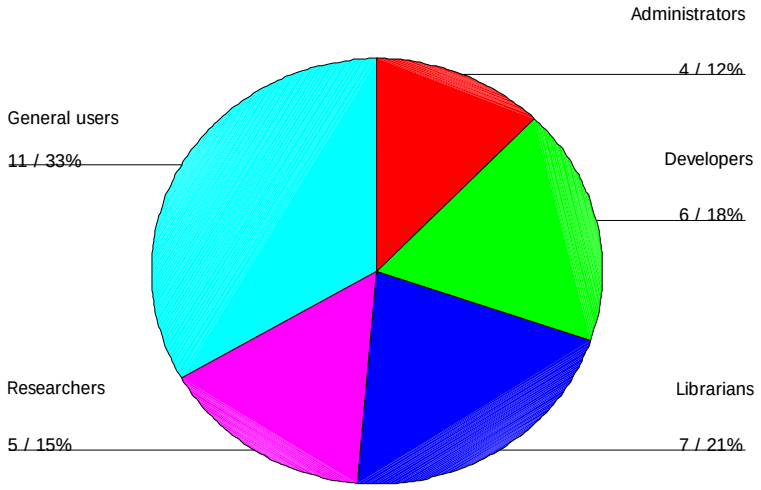

Figure 8.1: Participants' Stakeholder Group Distribution

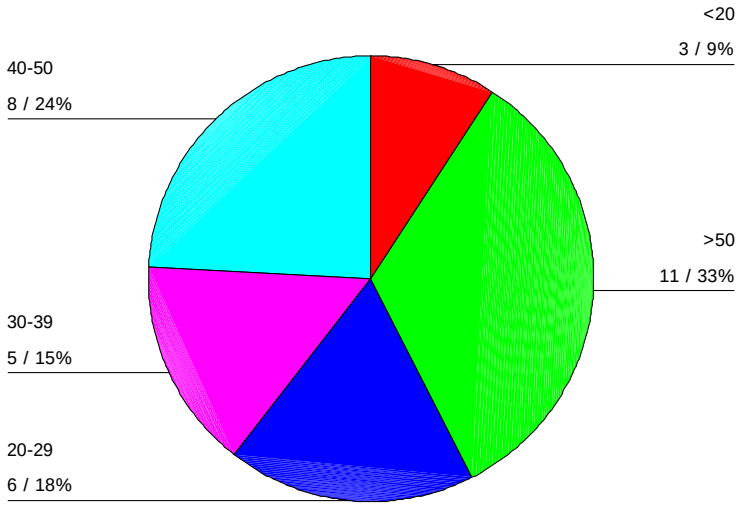

Figure 8.2: Participants' Age Distribution in Years

More than half of the participants (19, 57%) were over 40 years old, of which 11 (33%) participants were in their 50s. Additionally, 5 (15%) participants were in their 30s, 6 (18%) in their 20s, and 3 (9%) were under 20. The participants under 20 years old were all undergraduate users, whereas the ones over 50 were all DL administrators.

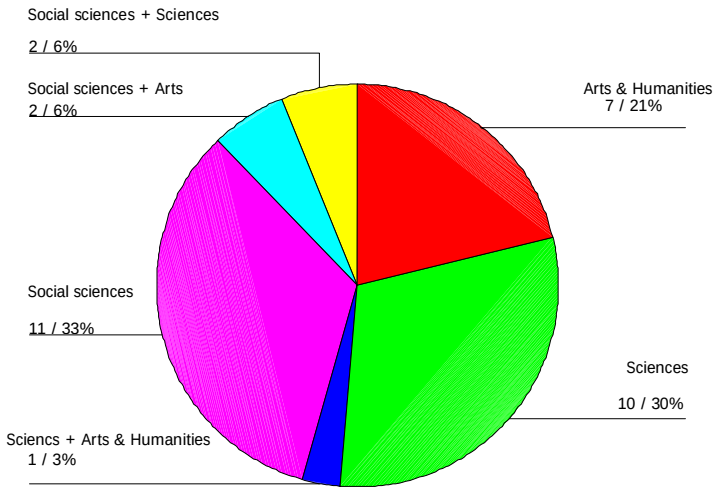

Figure 8.3: Participants' Subject Area Distribution

The composition of subject fields for these participants was 11 (33%) for social sciences, 10 (30%) for sciences, and 7 (21%) for humanities and the arts. Additionally, 5 (15%) participants reported two interdisciplinary fields (e.g., social science plus art & humanities). More than half of the participants (9, 58%) had been searching the RUL Web for books, journal articles, and other materials for more than three years, and over three-fourths (25, 76%) of the participants used the Web on a daily to weekly basis. Only one freshman was a first-time user of the RUL Web.

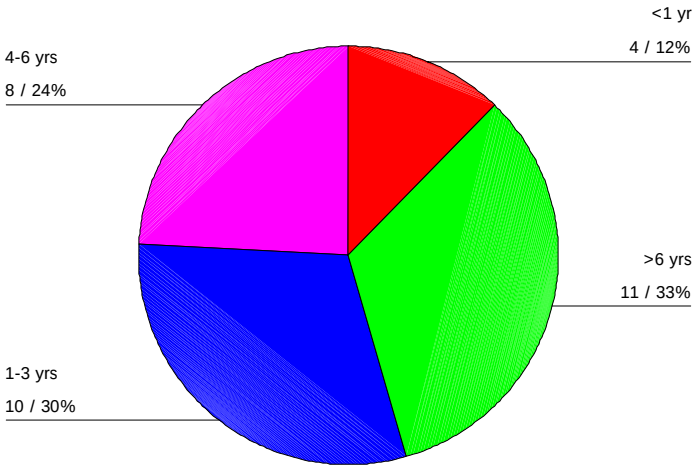

Figure 8.4: Participants' RUL Web Searching Distribution in Years

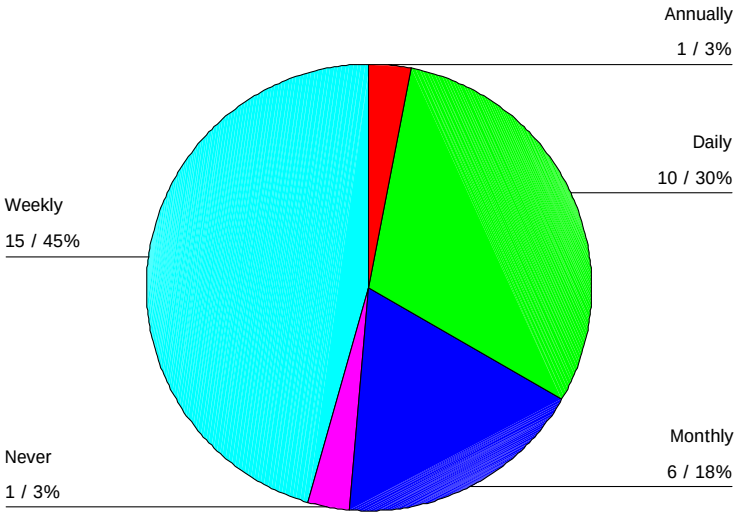

Figure 8.5: Participant RUL Web Searching Frequency Distribution

8.2 The Participants' Search Tasks

Although the participants came up with their own search tasks of various topics, the tasks per se could be grouped into two types: searching books, articles, images, and other Web resources about a given topic, and locating known items (e.g., e-reserve articles/book chapters, a book with known titles/authors). There was little difference among the stakeholder groups in terms of the types of search tasks. In other words, no matter whether they were DL administrators, developers, researchers, librarians or general users, their search tasks were essentially either books/articles/images/other Web resources about a given topic (26 cases, 79%) or locating known items (7 cases, 21%). The outcomes deviated from the researcher's expectations, which was that different groups DL stakeholders might vary in their interests in a DL.

8.3 The "Don't Know" Answers

The participants' "*don't know*" answers were treated as missing data, because they had no impact on the importance of the evaluation criteria. Rather, the answers could perhaps mean that the survey participants either didn't figure out whether a given criterion was importance or not, or they just simply didn't understand the criterion. Among the total amount of 39 criteria, 20 criteria (51%) did not receive any "don't know" answers. The other 19 criteria had more or fewer "don't know" answers, ranging from 1 (i.e., *usefulness* of information to target users*,* interface's *supportiveness to HCI, effort needed, the responsiveness* of digital services*, and copyright compliance*) to 7 (i.e., users' *behavior change*). In general, the criteria at interface (25%), service (40%) and technology (33%) levels received a smaller proportion of the "don't know" answers in contrast to the ones at content (56%), user (83%) and context (75%) levels. The findings are similar to the ones in the earlier confirmation stage.

8.4 The "Not Applicable to My Case" Answers

When asked to assign a importance rating to each criterion, the participants also could select more appropriate answers (i.e., "don't know" or "not applicable to my case") in relation to their searching experience with the RUL Web site for their tasks. Eventually, among the 39 criteria included in the experiment, 31 (80%) received the "not applicable" answers with total frequencies ranging from one to eleven. In general, the interface level had the most (3) criteria without the "not applicable" answers, whereas none of the content and technology level evaluation criteria received zero "not applicable" cases. The eight criteria without the "not applicable" answers are e*ffectiveness, consistency* and *ease of use* of the DL interfaces, users' *efficiency of task*

completion and *acceptance, technological efficiency, service reliability* and *sustainability.* Additionally, six out of the eight criteria were the core criteria (those in bold) from the proposed model, whereas the other three were the case-by-case criteria.

Meanwhile, eight criteria had more than five "not applicable" cases. They are service *courtesy (11* cases*), responsiveness (10* cases*), integrity to search path (5* cases*),* and *usefulness to target users (5* cases*), technological flexibility (10* cases*),* users' *productivity (6* cases*), extended social effect (7* cases*)* and *copyright compliance (5* cases*).* Since most participants did not use additional online service functions, it is not surprisingly that service level criteria had the highest counts of "not applicable" cases.

8.5 Participants' Criteria for DL Evaluation

At the beginning of the post-search questionnaire, the participants were asked to provide features/functions of the RUL Web that either had assisted or hindered their searching. The features they described were mostly related to digital content, technology, interface, and service. The participants seldom mentioned features/functions at the user and context levels. The only except was the *network effect,* to which two participants mentioned incoming and outgoing hyperlinks. The findings agree with the combined top ten most important criteria in the earlier section.

With very few new criteria, the features/functions indicated by the participants could be grouped into the existing 37 criteria in the proposed model. In other words, the experiment has verified that the proposed DL evaluation model comprised essential criteria for the RUL Web use setting. Table 8.1 below provides a small sample of quotes representing the existing criteria. They are the participants' criteria for DL evaluation in their own words.

Table 8.1: Participants' Representative Criteria for Digital Library Evaluation

Criteria	Representative quotes
CT-Comprehensiveness	"Some important journals are missing"; "Not enough books on the topic" "limited resources of databases in some areas"
CT-Accessibility	"Having the articles online is very convenient"
CT-Integrity of information	"PubMed access in addition to OVID Medline was useful as needed sequence info"
CT-Ease of understanding	"'Search IRIS'? confusing, many people don't know. What IRIS means? Probably better if just use 'catalogs'"
CT-Usefulness	"Given specific details of location" , "notification of availability" (were useful)
IF-Ease of use	"Not clear how to do phrase search although quoted phrases seems to work"
IF-Clarity	"'Search IRIS & Other Catalogs' or 'Find Articles'. What's the difference?"; "Clean display of IRIS record"; "Books are mixed up with journals in my Psycinfo search + users can't tell the difference" ; "Clear links"
IF-Consistency	"Non-availability of top bar at the front page (Like my account)"
IF-Appropriateness to target users	"Indexes + database list--if naïve user would not know which publisher 'packages' e.g. ScienceDirect, would be useful for searching"
IF-Personalization	"ACM-ability to change ordering from 'relevance' to 'date'"; "Results can't be sorted"
TN-Interoperability	"No federated search. I need to g to each journal or collection repeatedly and conduct my search"; "No easy way to switch from searching IRIS to searching databases (such as EBSCOHost)"; "'Finding articles (full texts)' from the results of the databases don't provide easy access to print resources. One has to search by ISSN number until users which users would not pay attention during the search at first"
TN-Accessibility	"Easy Net ID login (I worked from home)"
TN-Effectiveness	"Unable to see vernacular script", "Unable to search by vernacular script"; "Titles do not always contain the search phrase (in the search by title)"
TN-Efficiency	"The details page took a while to come up"
TN-Reliability	"Link resolver doesn't always work --one of my results went to a blank screen."
SV-Accessibility	"Was able to send it (results) to my email account"; "electronic document like the reserve have users' guide"; "You can ask questions online"; "the page suggested search words I could use"; "neither of these databases displayed 'help' for searching"
CX-Network effect	"Rutgers libraries holdings available via Google Scholar"; "Proper exterior links"

Meanwhile, there were a few criteria that were not included in the proposed model but were mentioned by the participants. They are:

- *CT-diversity* – "have several different databases available"
- *CT*-adequacy—"not sufficient information about a book on the direct results page"

- *CT-timeliness*—"Some of the literature to my topic was old and probably outdated"
- *CT-ease of navigation*— "On the first page, choosing where to go is difficult and ambiguous"
- *IF-error handling* — "Small but annoying-must 'Submit' in IRIS search (rather than ENTER)", "If no matches are found, the button to reverse the search does not work"

8.6 Participants' Most and Least Important Criteria at DL Levels

Table 8.2 through Table 8.7 below summarize the top five most important criteria and the least regarded criterion (in italics) respectively for the evaluation at the six DL levels perceived by the 33 participants. Meanwhile, comparisons have been made in the right columns between these criteria and those highly regarded by the 431 survey participants in the confirmation stage. The rankings of the importance ratings are based upon descriptive data (i.e. the mean scores and then the standard deviation). The larger the mean score and the smaller the standard deviation, the higher the ranking.

In general, there is a high replicability between the two stages (confirmation and verification) regarding the important criteria for DL evaluation at the six levels. Among 27 top ranked criteria from the experiment, more than three fourths (21, 78%) are also on the top lists in the survey (see those in Table 8.2 through Table 8.7 with "x" in the right columns). Four of the remaining six criteria (*CT-ease of understanding, TN-security, IF-supportiveness to HCI,* and *UR-acceptance*), although not on the top five important criteria from the survey, were all ranked as sixth among about eight or nine criteria at the corresponding DL levels. Additionally, except for the least criterion for context

evaluation, all of the other five least perceived criteria remained identical on both the experiment and survey lists. Furthermore, digital service criteria received the highest consistency between the survey and the experiment (all top five criteria and the least criterion remained the same between the two stages, and also the same in the exploratory stage), whereas the context level had the most divergence in important and unimportant criteria.

8.6.1 Content Level Evaluation

As for the digital content evaluation, *usefulness to target users* remained the top criterion, which also appeared earlier in the top five lists of the survey, the interview card sorting (CS), and the interview open question answering (QA). *Accessibility, ease of understanding, accuracy* and *comprehensiveness* were the other top criteria. In particular, *ease of understanding* was the sixth in the survey, where the administrator survey participants tended to provide lower ratings. However, in searching a DL (RUL Web) for their own tasks (mostly finding books, book journals, images), the participants tended to increase the weight of the criterion, especially when they failed to locate the information they needed. In contrast, *fidelity* and *integrity of information* from the top five list of the survey dropped to the eighth and seventh out of nine in the experiment, presumably because most searching tasks were to find books/articles for given simple tasks, and there was less need for the two criteria.

Unanimously, the experiment and the survey participants regarded *conciseness of information* as the least important criterion for digital content evaluation.

Table 8.2: Participants' Top Five and the Least Important Criteria/Criterion (Content)

Criteria	Top five & the least important criterion in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Usefulness	5.67 (.61)	x
Accessibility	5.63 (.61)	x
Ease of understanding*	5.58 (.96)	
Accuracy	5.37 (1.03)	x
Comprehensiveness**	5.33 (.88)	
Conciseness	*4.45 (1.53)*	*x*

*Was 6th out of the nine content evaluation criteria;
**Was 8th out of nine content evaluation criteria

8.6.2 Technology Level Evaluation

Digital technology evaluation also achieved a high consistency between the survey and the experiment results. Technological *efficiency*, *reliability, east of use* and *interoperability* among systems appeared in the top lists from both stages. *Security* also had a high mean score (6.06) in the survey, which should be scaled above "somewhat significant". Again, *flexibility* was considered as the least important criterion by both survey and experiment participants.

Table 8.3: Participants' Top Five and the Least Important Criteria/Criterion (Technology)

Criteria	Top five & the least important criterion in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Efficiency	5.66 (.60)	x
Reliability	5.63 (.89)	x
Security*	5.55 (.99)	
Ease of use	5.52 (.89)	x
Interoperability	5.25 (.97)	x
Flexibility	*5.05 (1.02)*	*x*

*Was 6th out of the nine technology evaluation criteria

8.6.3 Interface Level Evaluation

Interestingly, when people do real searching on an operational DL system, their perceived importance for the interface's *supportiveness of HCI* increased. Additionally, they consistently perceive *effectiveness, ease of use, effort needed* and *consistency* of a digital interface to be important for their searching tasks. *Appropriateness to target users*

dropped to sixth in the experiment, probably because all participants were targeted users of the RUL Web site, and almost all used the site on a daily to weekly basis. Hence, there should not be an issue regarding whether or not the RUL Web fits is part of the participants' background. *Attractiveness* is still the second lowest ranked criterion in the experiment, and *personalization* remains the least importance.

Table 8.4: Participants' Top Five and the Least Important Criteria/Criterion (Interface)

Criteria	Top five & the least important criterion in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Effectiveness	5.79 (.48)	x
Ease of use	5.58 (.66)	x
Supportiveness of HCI	5.23 (1.02)	Was the 6th out of eight
Effort needed	4.74 (1.41)	x
Consistency	4.67 (1.43)	x
Personalization	*3.07 (1.62)*	*x*

8.6.4 Service Evaluation

Similar to the survey participants, the experiment participants viewed *reliability, usefulness, responsiveness* and *integrity to information seeking path* as the most important criteria at the service level, while they assigned the least weight to the *courtesy* of a digital service. Up to date, digital service evaluation criteria have been achieving the highest consistency across the research stages, including the exploratory (interview), the confirmation (online survey), as well as the verification (experiment).

Table 8.5: Participants' Top and the Least Important Criteria/Criterion (Service)*

Criteria	Top five & the least important criterion in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Reliability	5.57 (.63)	x
Usefulness to target users	5.29 (1.01)	x
Responsiveness	5.23 (1.15)	x
Integrity to information seeking path	4.75 (1.24)	x
Courtesy	*4.30 (1.45)*	*x*

*Accessibility was not in the post-search questionnaire of the experiment—this was a flaw in the experiment design

8.6.5 User Level Evaluation

The participants' perceived important criteria for user level DL evaluation were consistent with those from the proposed DL evaluation model. Both survey and experiment participants regarded *successfulness* and *efficiency of task completion, use/reuse* as well as *satisfaction* as the most important criteria. Even users' *acceptance*, which was not on the top-five important criteria list, was sixth out of the nine criteria in the survey.

Table 8.6: Participants' Top Five and the Least Important Criteria/Criterion (User)

Criteria	Top five & the least important criterion in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Successfulness of task completion	5.72 (.81)	x
Efficiency of task completion	4.91 (1.01)	x
Use/reuse	4.84 (.85)	x
Acceptance*	4.65 (.98)	
Satisfaction	4.47 (1.11)	x
*Productivity***	*3.82 (1.50)*	

*Was 6th out of the nine user evaluation criteria
**Was 5th out of the nine user evaluation criteria

In contrast to the consensus on the most important criteria, the least important criterion had divergence between the two research stages. Dramatically, the fifth criterion (*productivity*) in the survey dropped down to the least important criterion in the experiment. The original least criterion (*behavior changes)* became the second least for evaluating a DL at the user level with mean=4.17, SD=1.44. Presumably, *productivity* had the least connection with their searching experience on the RUL Web for their tasks during the experiment. This echoes the survey findings that DL users tend to care more about direct effects (e.g., *efficiency* and *successfulness of task completion*) over indirect effects (e.g., *behavior changes, productivity*) of using a DL. However, considering the

large standard deviations (1.5 for productivity and 1.44 for behavior changes), it might need further investigation on the lowest ranked results.

8.6.6 Context Level Evaluation

Similar to the DL evaluation at the user level, context evaluation received less agreement between the survey and the experiment participants, in particular in the aspects of *network effect* and *extended social effect.* A DL's *extended social effect,* including supporting multi-disciplinary research, improving social and economic status of prospective users, preserving knowledge and culture, etc., rose from the least criterion in the survey to the top of the list in the experiment, which is similar to the earlier interview findings. Meanwhile, *network effect* dropped from the top-five list in the survey to the least important criteria in experiment. It might be worthwhile to explore the underlying reasons, but no specific reason can be given at this time. No other data could be found to support the findings. Despite the divergence, *sustainability* and *copyright compliance* have consistently received the highest rankings across the research stages for DL context evaluation.

Table 8.7: Participants' Top and the Least Important Criteria/Criterion (Context)*

Criteria	Top five & the least important criteria in the experiment with mean (SD) n=33	Appearance in the survey top five and the least list
Sustainability	5.48 (.76)	x
Copyright compliance	4.96 (1.51)	x
Extended social effect**	4.43 (1.41)	
*Network effect****	*3.81 (1.42)*	

*Only four contextual criteria were included in the post-search questionnaire of the experiment. Collaboration/sharing and Managerial support were deliberately excluded from the post-search questionnaire with a consideration of not applying into the RUL searching setting.
**Was the least among the nine user evaluation criteria in the survey;
***Was 5th out of the nine content evaluation criteria in the survey

8.7 The Combined Most and Least Important Criteria at the Six DL Levels

Table 8.8 and Table 8.9 show a mixed list of the top important criteria and a mixed list of the least important criteria across the six DL evaluation levels. Similar to the findings of the confirmation stage, the combined top ten criteria are composed of more criteria for lower DL level evaluation (i.e., content, technology, interface, and service) as opposed to higher-level evaluation (i.e., user and context). Specifically, the top 10 list is composed of three content criteria, three technology criteria, two interface criteria, one service criterion and one user criterion. In contrast, the least list contains more higher-level evaluation criteria, including two context criteria, three user criteria along with three interface criteria, one service criterion, and one content criterion. Not surprisingly, when searching a DL with their tasks, the stakeholders were more concerned about the DL aspects with which they are directly interacting, including content, technology, interface, and service.

Table 8.8: The Top Ten Criteria for the Six DL Levels (n=33)

Rank	Criteria	Mean	SD
1	IF-effectiveness	5.79	0.48
2	UR-successfulness of task completion	5.72	0.81
3	CT-usefulness	5.67	0.61
4	TN-efficiency	5.66	0.60
5	TN-reliability	5.63	0.89
6	CT-accessibility	5.63	0.61
7	CT-ease of understanding	5.58	0.96
8	IF-ease of use	5.58	0.66
9	SV-reliability	5.57	0.63
10	TN-security	5.55	0.99

Again, by comparing the top ten mixed criteria included in the experiment and the earlier survey, one finds consistency between the lists. Six criteria (60%) from the survey top-ten list have also been on the experiment top-ten list. They are content *accessibility,*

technological reliability, effectiveness and *ease of use* of interface, *service reliability,* and *users' successfulness of task completion*. Additionally, three out of the other four criteria (i.e., *technological ease of use, sustainability,* and *information accuracy*) are ranked as 11th, 12th, and 13th respectively in the experiment list. Similarly, six criteria (60%) from the survey's least-ten list (i.e., *interface personalization aesthetically attractiveness,* users' *behavior change, conciseness of information, extended social impact,* and *service courtesy*) were on the least-ten list of the experiment as well. Both survey and experiment participants ranked them as the least important for DL evaluation.

Table 8.9: The Least Ten Criteria for the Six DL Levels (n=33)

Rank	Criteria	Mean	SD
1	IF-personalization	3.07	1.62
2	CX-network effect	3.81	1.42
3	UR-productivity	3.82	1.50
4	UR-behavior changes	4.17	1.44
5	IF-attractiveness	4.26	1.63
6	SV-courtesy	4.30	1.45
7	IF-appropriateness to target users	4.38	1.43
8	CX-extended effect	4.43	1.41
9	CT-Conciseness	4.45	1.53
10	UR-satisfaction	4.47	1.11

8.8 Consensus/Differences Among Stakeholder Groups

Unlike the earlier findings regarding the top and least importance rankings in which a fair consistency exists between the survey and the experiment results, a one-way ANOVA test shows that the group difference in the experiment was not consistent with the corresponding survey findings. Whereas 11 out of 51 criteria had a statistically significant inter-group divergence in the survey, only one criterion (*comprehensiveness* of digital content, one out of the 37 criteria in the survey) kept the group difference, with $F(4, 25)=6.174$, $p<.001$. Further, for the criterion, while the difference existed only

between librarians and users in the survey, in the experiment the group difference was switched to developers and three other stakeholder groups (librarians, researchers, and users). The previous librarian-user differences did not hold in the verification stage. Presumably, this is because in the experiment setting, in spite of the difference of stakeholder group affiliation, the participants were all searching the RUL Web for similar tasks, such as finding books, journal articles, images. The similar searching task patterns might have an impact on identifying true group differences.

Although only one criterion (content *comprehensiveness*) had a statistically significant difference among the stakeholder groups, some potential divergences in the top perceived important criteria among the five stakeholder groups were detected at each of the six DL levels (see Table 8.10 through Table 8.15). The tables show that for a given DL level evaluation, some criteria are on the top five lists of all stakeholder groups (core evaluation criteria, and those in bold), while the others are perceived as being important only by some of the groups (case-by-case evaluation criteria). The following sections will discuss what consensus and divergences are identified from the experiment regarding highly-perceived important DL evaluation criteria.

8.8.1 Content Level Evaluation

All the five stakeholder groups regarded *accessibility* to digital information/collection, *accuracy* of information and *usefulness* to target users as very important to digital content evaluation.

Meanwhile, except for the researcher group, the other four groups agreed that *ease of understanding* should be an important criterion as well. As the only criterion with statistically proven inter-group divergence, *comprehensiveness* of a digital collection was

highly perceived by the librarians, researchers and general users but was absent from the administrators' and developers' top five lists. Additionally, almost all groups except the general users have their own unique important criteria that were not on the remaining groups' top five lists. They are *appropriateness to target users* from the librarian group, *fidelity of information* from the administrator group, *integrity of information/collection* from the developers, and *conciseness* of information from the researcher group. Interestingly, *conciseness* of information used to be the least important criterion in the survey (out of nine criteria). But in the experiment, it jumped to fourth out of the nine criteria. The reason for this is not clear.

Table 8.10: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (Content)

Criteria	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Accessibility	**X**	**X**	**X**	**X**	**X**
Accuracy of information	**X**	**X**	**X**	**X**	**X**
Usefulness to target users	**X**	**X**	**X**	**X**	**X**
Ease of understanding	X	X	X		X
Appropriateness to target users			X		
*Comprehensiveness			X	X	X
Fidelity	X				
Integrity of information		X			
Conciseness of information				X	

***** The one that was statistically proven to have the inter-group difference.

8.8.2 Technology Level Evaluation

For assessing digital technology, on the one hand, four criteria received inter-group consensus: *ease of use, reliability, security,* and *efficiency.* On the other hand, two criteria (i.e., *interoperability* and *flexibility*) had inter-group divergence. Whereas the administrators, librarians and general users were more concerned about *interoperability* across systems, the developer and researcher groups gave greater weight to technological

flexibility, which was the least (ninth) criterion for the researchers and seventh out of nine criteria for the developers in the online survey.

Table 8.11: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (Technology)

	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Ease of use	**X**	**X**	**X**	**X**	**X**
Reliability	**X**	**X**	**X**	**X**	**X**
Interoperability	X		X		X
Security	**X**	**X**	**X**	**X**	**X**
Efficiency	**X**	**X**	**X**	**X**	**X**
Flexibility		X		X	

8.8.3 Interface Level Evaluation

There were eight interface level criteria in the experiment. Four received inter-group consensus. All five groups agreed that *ease of use, effectiveness, consistency,* and *supportiveness to HCI* were important criteria for assessing DLs at the interface level. Additionally, *effort needed* was also perceived to be an important criterion by four out the five groups, except the librarian participants. When searching the Rutgers Library Web site for their own tasks, only the librarian groups still believed that a DL interface should be constructed to be appropriate to target users' background and needs.

Table 8.12: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (Interface)

	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Ease of use	**X**	**X**	**X**	**X**	**X**
Effectiveness	**X**	**X**	**X**	**X**	**X**
Consistency	**X**	**X**	**X**	**X**	**X**
Appropriateness to target users			X		
Supportiveness of HCI	**X**	**X**	**X**	**X**	**X**
Effort needed	X	X		X	X

8.8.4. Service Level Evaluation

All five stakeholder groups agreed on the important criteria for digital service evaluation. In their view, it is important for a given digital service or service agent to be

reliable, responsive, useful to target users, and should be well *integrated to information seeking path. Courtesy* again was the least criterion. *Service accessibility*, one of the important criteria from the previous stages, was accidentally omitted in the experiment post-search questionnaire, and thus offered no data.

Table 8.13: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (Service)

	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Integrity to information seeking path	**X**	**X**	**X**	**X**	**X**
Reliability	**X**	**X**	**X**	**X**	**X**
Responsiveness	**X**	**X**	**X**	**X**	**X**
Usefulness to target users	**X**	**X**	**X**	**X**	**X**

8.8.5 User Level Evaluation

Among seven user level criteria included in the post-search questionnaire, three were unanimously perceived by the five stakeholder groups as important to DL evaluations: *successfulness and efficiency of task completion,* as well as *use/reuse.* Satisfaction, one of the core evaluation criteria in the preceding stage (i.e., the online survey) dropped out of the core list in the experiment, because it was not on the top list of the administrator group. However, considering the very small sample size of the group, the result might not reflect a real population situation. Accordingly, more verification experiments need to be conducted.

In contrast to the four criteria that were widely perceived to be important among the groups, the remaining three criteria received diverse rankings from the stakeholders. For instance, *acceptance* was included in the top lists of three of the five groups (i.e., administrator, developer, and user), but was excluded from the librarians' and researchers' top lists. Similarly, *productivity* was perceived to be an important indicator only by the librarian group but not the other four groups. Interestingly, *behavior change*

was highly ranked by the administrator and researcher participants, but it was the least criterion in the online survey.

Table 8.14: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (User)

	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Successfulness of task completion	**X**	**X**	**X**	**X**	**X**
Satisfaction		X	X	X	X
Efficiency of task completion	**X**	**X**	**X**	**X**	**X**
Use/reuse	**X**	**X**	**X**	**X**	**X**
Acceptance	X	X			X
Productivity			X		
Behavior change	X			X	

8.8.6 Context Level Evaluation

For the context level evaluation, two criteria received the same importance ratings among the five stakeholder groups: *sustainability* and *copyright compliance*. Meanwhile, *extended social effect* was also widely perceived to be an important indicator by four out of the five groups, except the researcher participants. Ironically, in the survey setting, the researcher group was the only one who had this criterion in its top five lists.

The post-survey design for the context level evaluation criteria is problematic, because only four criteria were covered due the consideration of non-applicability of other criteria (e.g., *collaboration, managerial support, integrity to social practice, productivity of community members, outcome against organizational goal*s) in a university library Web site search setting. Consequently, the findings might not be valid. Therefore, a revised verification test with more context level evaluation criteria included is needed.

Table 8.15: Comparison of the Top Five Criteria Among the Five Stakeholder Groups (Context)

	Administrators (n=4)	Developers (n=6)	Librarians (n=7)	Researchers (n=5)	Users (n=11)
Sustainability	**X**	**X**	**X**	**X**	**X**
Copyright compliance	**X**	**X**	**X**	**X**	**X**
Extended social effect	X	X	X		X
Network effect				X	

8.9 Verification of the Proposed DL Evaluation Model

By comparing these tables with their counterparts from the survey stage, there is a fair consistency between the two sets in terms of (1) important criteria for DL evaluation at all six levels, and (2) important criteria's core and case-by-case affiliations.

First, except for the four criteria (i.e., *service accessibility, technology effectiveness, collaboration, and managerial support*) that were excluded in the experiment post-questionnaire either by mistake or on purpose, all important criteria in the proposed DL evaluation model from the confirmation stage were on the top five lists of the stakeholder groups. In particular, the important criteria for interface (i.e., *ease of use, effectiveness, consistency, interaction support, effort needed* and *appropriateness*) and service evaluation (i.e., *reliability, responsiveness, usefulness* and *integrity*) were exactly the same between the two research stages. For content, technology and user level evaluations, there was only one new important criterion for each. Further, each of the three new criteria was on top five lists of very limited stakeholder groups. For instance, content *conciseness* was only on the researcher group's list, technological *flexibility* was on the researchers' and developers' list, and *behavior change* was only on the researchers' and administrators' lists.

Second, the high percentage (81%) of the criteria in the proposed model has been placed correctly as either core or case-by-case criteria according to the experiment

findings. Only 6 out of 37 criteria might be placed in the wrong categories, among which one core criterion (user's *satisfaction*) might be case sensitive and five case-by-case criteria (technological *security and efficiency; supportiveness of HCI, use/reuse,* and *copyright compliance*) might need to be changed to core criteria. The consistency seems to be more valid in the core criteria part of the proposed DL evaluation criteria. Eighteen out of the 19 (95%) core criteria have received consistent results in the survey and the experiment.

However, as for the case-by-case criteria, although the suggested case-by-case criteria from the experiment are essentially the same as those from the confirmation stage (i.e., the criteria in the outer rings of the graphic DL evaluation model), some changes have been observed regarding which stakeholder groups perceived which criteria to be important. The original big divergence between the general user group and the other four groups did not hold in the experiment settings. Similarly, no more observations could be made to support earlier findings regarding the frequently shared perspectives among the three groups from the librarianship domain (i.e., administrators, developers and librarians).

The following sections will discuss in detail one DL level-by-level regarding how the important criteria (both core and case-by-case) in the proposed model are either supported or rejected by the experiment findings.

8.9.1 Content Level Evaluation Criteria

For digital content evaluation, the experiment suggested the same three core criteria (*accessibility, accuracy,* and *usefulnes*s) as depicted in the proposed model. As for case-by-case criteria, *ease of understanding* was still widely perceived to be important

by four out of the five groups. However, the composition of the four groups had a slight change. While the other three groups remained the same in both stages, it was the administrator, not the researcher group, in the experiment that perceived *ease of understanding* as one of the top five. In addition to general users, in the experiment setting, *collection comprehensiveness* was also perceived to be important by two other groups (i.e., librarians and researchers). Further, it is the only case-by-case criterion for which the inter-group difference has been statistically proven in the experiment setting. In contrast, the number of groups favoring *appropriateness to target users* dropped from three (i.e., researchers, librarians and administrators) in the survey to only one (librarians) in the experiment. Integrity and fidelity of information switched the favored groups between the administrators and the developers.

8.9.2 Technology Level Evaluation Criteria

When searching an operational DL for their tasks, there were more agreed-upon core evaluation criteria among DL stakeholder groups for technology evaluations. In additional to the two criteria (i.e., *reliability* and *ease of use*) suggested in the proposed model, the experiment participants regardless of their DL roles unanimously regarded *efficiency* and *security* to be one of the important criteria. As for case-by-case criteria, the developers and researchers no longer ranked *interoperability* as one of the top five. Instead, both groups thought technological flexibility might be more important. Meanwhile, general users became more interested in the criterion.

8.9.3 Interface Level Evaluation Criteria

The three core criteria (i.e., *ease of use, effectiveness,* and *consistency*) in the proposed model were well supported by the experiment findings. In addition, the

experiment suggested that supportiveness of HCI could be upgraded as a core criterion for digital interface evaluation. Upon checking back on the survey results, the criterion was categorized as case-by-case only because it was ranked as sixth by the researcher survey participants while being in the top 5 on the other groups' lists. As such, the suggestion likely is valid.

As for the remaining case-by-case criteria, two more groups (i.e., administrator and developer) added *effort needed* to their top five lists in addition to the original user and researcher groups. Perhaps, when the stakeholder groups conducted real searches on the library Web site, *effort needed* became more important to them. In contrast, the number of groups favoring *appropriateness to target users* dropped from four (i.e., administrators, developers, librarians, and researchers) in the survey setting to one (librarian) in the experiment setting.

8.9.4 Service Level Evaluation Criteria

Service criteria again received the greatest inter-group consensus in the experiment. There were no case-by-case criteria. All core criteria except accessibility (omitted by mistake in the post-search questionnaire) were also verified to be core in the experiment: *reliability, responsiveness, usefulness to target users,* and *integrity to information seeking path.*

8.9.5 User Level Evaluation Criteria

Different from the other level evaluation whereby core evaluation criteria were highly stable across the two research stages, two plausible core criteria changes are implied. Users' *satisfaction* was not in the administrators' top five list. Accordingly, it might need to be downgraded as a case-by-case criterion. However, its validity is

questioned by the very small size of the administrator sample. In other words, further verification needs to be made before any action is taken. Meanwhile, according to the experiment findings, use/reuse was suggested to be upgraded as a core criterion.

Other than the two criteria, the rest of the user level core criteria (i.e., *successfulness* and *efficiency of task completion*) remain the same. Nevertheless, the association of the case-by-case criteria as to the stakeholder groups had some changes in the experiment. First, *behavior change* was added as a new case-by-case criterion, because it was indicated by both administrator and researcher participants. Meanwhile, the number of groups that perceived *acceptance* as important criterion dropped from four (administrators, developers, librarians, and researchers) in the survey to three (administrators, developers, and general users) in the experiment. Additionally, *productivity* was no longer perceived to be important by the general users and researcher groups in the experiment setting. Instead, the librarians were the only group who perceived this criterion to be important.

8.9.6 Context Level Evaluation Criteria

As mentioned earlier, there was a flaw in the post-search questionnaire design for context level evaluation criteria verification. Only four criteria were included in the experiment. As such hardly any valid verification can be made on the results. Nevertheless, it is true that all the five groups assigned higher rankings to *sustainability* and *copyright compliance* than to *extended social effect* and *network effect.*

In sum, fairly highly consistent results were developed across the research stages regarding what criteria should be important, including core and case-by-case, for DL evaluations. However, there are inconsistent results for group association for some case-

by-case criteria. Since no other criteria except content *comprehensiveness* was proven to have a statistically significant inter-group divergence, the verification of the case-by-case criteria part of the model cannot be made via this experiment. Further, in spite of the plausible change indication from the experiment, no modifications should be made in the model for the time being, considering the following defects in the experiment:

- Except comprehensiveness of collection, none of other case-by-case criteria in the model has been statistically proven between-stakeholder-group difference in the experiment. Presumably, this is because in the experiment setting, in spite of the difference of stakeholder group affiliation, the participants were all searching the RUL Web for their tasks with similar patterns, such as finding books, journal articles, images. The similar searching task patterns might influence the identification of group differences. Therefore, further experiment verification should consider using standardized tasks across participants;
- Users' satisfaction was not on the participating administrator group's top list in the experiment, and thus its core evaluation status is questionable. However, considering the inadequate sample size, especially for administrators (n=4), the experiment finding might not be valid. In the future, more verification experiments with larger samples size should be conducted.
- The inconsistent results at the context level might owe to the incomplete inclusion of all context evaluation criteria in the model to the post-search questionnaire. For instance, *managerial support*, one of the core context evaluation criteria, was deliberately excluded from the post-search questionnaire simply for its inapplicability to search settings. The exclusion might have a negative impact on the context level evaluation ratings and rankings in the experiment. Therefore, future experiments should consider including all criteria in the model.
- Except for *supportiveness of HCI*, the remaining criteria category changes based upon the experiment results cannot be supported by the earlier survey data.

The model needs to be tested in more DL use settings with more stakeholders' involvement and benchmarking representative tasks, while the design drawbacks need to be removed. Further, based upon the consistent results between the two stages regarding the top/least perceived criteria at each DL level and among DL groups, it can be claimed with confidence that the proposed holistic DL evaluation model has been verified by the experiment, especially at the content, interface and service levels.

Chapter 9 FURTHER DISCUSSION AND CONCLUSION

The dissertation research has been proved to be effective and significant from the following aspects: (1) consistently perceived important criteria across the three research stages; (2) proven inter-group divergence on criteria importance perceptions, especially the divergence between the user and the other fours groups from the academic and professional DL domains; (3) newly identified evaluation criteria augment the existing body of research; and finally (4) the construction of the holistic DL evaluation model. The validity of these findings have been strengthened through employing various complementary research techniques, embracing perspectives from diverse DL stakeholder groups, as well as acknowledging the promising framework proposed by DL experts. In spite of a few weaknesses in the research design, the research findings are valuable for further DL innovations. This chapter will start with further discussions on integrated findings across the three research stages with an emphasis on the implications for DL academic and professional domains. The discussion will then be followed by a summary of research strengths and weaknesses for which further studies are suggested.

9.1 Integrated Research Findings across the Three Research Stages

9.1.1 Consistently Perceived Important Criteria across the Research Stages

Through out the three research stages, many most/least perceived important criteria have been consistently identified for DL evaluations at the six levels. Table 9.1 shows the consistent top and least important criteria across the stages.

In general, what these DL stakeholders are concerned about are being able to access high quality content and service (**the Premise;** e.g., content and service *accessibility, sustainability*), Their second concern would be ease of search and use

during their interaction with the content and service (**the Process;** e.g., *ease of use, effort-needed, interoperability,* service *responsiveness***)**, and then, they would care about **the direct Performance** of using the DL, such as *usefulness, efficiency* and *successfulness of task completion*. In contrast, the least perceived criteria are those indirect outcomes of DL use (i.e., not directly related to finding expected information, such as. *behavior change, extended social impact*), or non-core processes and premises, such as *personalizability of an interface, courtesy of a service, conciseness of information*, etc. The consistently perceived and ranked criteria lists serve as a reliable base for the construction of the holistic DL evaluation model.

Table 9.1: Consistently Perceived Most/Least Important Criteria Across Research Stages

	Content	Technology	Interface	Service	User	Context
The most	accessibility accuracy usefulness	ease of use interoperability efficiency reliability	ease of use consistency, effort needed effectiveness;	accessibility, integrity responsiveness reliability, usefulness	successfulness efficiency satisfaction, use/reuse;	sustainability copyright compliance
The least	conciseness	display quality	attractiveness personalization	courtesy	behavior change	extended social impact

Similarly, when searching a DL for their tasks, the stakeholders were consistently concerned about the DL aspects with which they were directly interacting: content, technology, interface, and service. In general, contextual factors received lower priorities in their perspectives except sustainability of a DL. A plausible reason might be related to the complexity of and people's familiarity with a given level. Comparatively, service level is fairly straightforward, while context has broader coverage in terms of institutional, social, cultural, economic, and legal sub-levels.

The rankings of a few criteria changed across the studies. For instance, the technology *security* issue was one of the least perceived criteria in the interview (the

exploration stage), but jumped to the top of the list in the experiment (the verification stage) and became sixth out of nine in the confirmation stage (the online survey). Users may have become more aware of the issue when they actually interacted with a DL. Similarly, *learning effect* was perceived as the most important criteria for user level evaluation by the interviewees, but dropped out of the top five lists in the succeeding two research stages (the confirmation and verification). This may be because that the interview did not include general users and librarians as interview participants, who tended to place their emphasis on the direct effects of DL use.

In general, consistently perceived important and unimportant criteria across the research stages suggest a need for prioritization in DL research and development. DL researchers and developers should first make DL contents, technologies, and services accessible. Then they should provide easy search/use of these contents, technologies, and services through digital interfaces. Meanwhile, it is important to improve direct performance of DL uses.

9.1.2 Proven Inter-group Divergence on the Criteria Importance Perceptions

The research consistently identifies a divergence among the stakeholder groups regarding what criteria should be used for DL evaluation. These results are gathered through the interview and the survey. For instance, in both stages of the research, service, interface and user evaluation criteria received greater consensus among the stakeholder groups regarding the importance ratings. In contrast, technology, context and content evaluation criteria received more divergent rankings among the groups. The underlying reason for the lowest agreement regarding the important technology evaluation criteria is

presumably associated with the unfamiliarity with details of DL technologies for the majority of the stakeholders except the developers.

Additionally, it should be noted that not all DL evaluation criteria but only a small portion of the criteria has a statistically significant difference among the five DL stakeholder groups. Instead, only a small portion of the criteria holds statistically significant group differences. This suggests the feasibility of conducting a general DL evaluation embracing multiple viewpoints from various stakeholders.

9.1.3 Important Criteria Perceived by Users

Further, the research (especially the confirmation stage) implies that the group differences exist mostly between general users and the other stakeholder groups. Unlike the other stakeholder groups' perspectives, all *appropriateness* criteria for the aspects of digital content, technology, and interface were not favored by the general users. Instead, digital library users are more concerned about *comprehensiveness of collection*, their *effort needed* for interacting with a DL, *productivity*, and *network effects* in terms of incoming/outgoing links from/to other resources. The findings inform DL research and professional domains. Considering that in reality the other than general user groups comprise the key players in DL innovation and there has been very little users' involvement in DL development according to the interview findings, the research outcome provides an alert to both DL academic and professional domains about the diverse opinions of DL end users.

9.1.4 New Evaluation Criteria Augmenting the Existing Research Body

The integrated research findings across the three stages suggest that the existing DL evaluation research embraces the important criteria highly perceived by the various

stakeholder groups. This is especially true for the content, interface, service and user level evaluations. By comparing the proposed holistic DL model with the earlier literature review findings, one sees that there are gaps between what should be evaluated and what have been evaluated. Table 9.2 lists the criteria suggested by the model that have been adopted in the previous studies.

From the table, among the 37 important criteria from the proposed model, including the core and the case-by-case criteria (the ones with parentheses within which the stakeholder groups opting for the criteria are indicated), 123 (32%) criteria have not yet been examined in any previous studies. The unexplored criteria are primarily from the context and technology levels. In contrast, all interface evaluation criteria have been adopted. For the context level evaluation, only *copyright compliance* has been investigated by the Human-Computer Interaction Group at Cornell University (Jones et al., 1999) when they investigated digital collection evaluation efforts across five different DL prototype projects. Although *sustainability*, as one of the core criterion for context level DL evaluation, was suggested earlier by Blixrud (2002) and Lynch (2003), so far no empirical evaluation studies have been found to address the issue. This again supports Saracevic's assertion (2000) that contextual effects of DL have not been adequately addressed. Technology level evaluation is another weak area. One core (i.e., *ease of use*) and two case-by-case (i.e., *interoperability* and *security*) criteria have not yet been used in any DL evaluation research, in spite of the suggestion in Kwak et al.'s DL evaluation framework (2002) for addressing the *security* issue.

Table 9.2: The Adoption Status of the Important Criteria in the Existing Studies

Criteria in the model	Existing evaluation studies with the criteria adopted
CT-accessibility	Adams & Blandford, 2001; Bishop, 1998; Jones et al., 1999; Wilson et al., 2002
CT-accuracy	Bergmark et al., 2002; Jones et al., 1999; Machionini et al., 2003; Zhang et al., 1995
CT-usefulness	Zhang et al., 2004
CT-appropriateness (ADM, LIB, RES)	Borgman et al., 2001; Ding et al., 1999
CT-comprehensiveness (USR)[1]	
CT-ease of understanding (DEV, LIB, RES, USR)	Khoo et al., 2002, Zhang et al., 2004
CT-fidelity (DEV)	Jones et al., 1999; Kenney et al., 1998
CT-integrity (ADM)	
TN-ease of use	
TN-reliability	Champeny et al., 2004; Papadakis et al., 2002
TN-effectiveness (ADM, LIB, RES, USR)	Bosman et al., 1998; Hee 1999; Jones & Lam-Adesina, 2002; Khoo et al., 1998; Larsen, 2000; Rui et al., 2000; Salampasis et al., 2002; Sanderson & Crestani, 1998
TN-efficiency (DEV, RES, USR)	Fuhr et al., 2002; Kapidakis et al., 1998; Kengeri et al., 1999; Larsen, 2000; Xi et al., 2002
TN-interoperability (ADM, DEV, LIB, RES)[2]	
TN-security (ADM, DEV, LIB)[3]	
IF-ease of use	Champeny et al., 2004; Hill et al., 1997; Huxley, 2002; Khoo et al., 2002; Papadakis et al., 2002
IF-effectiveness	Browne, 2001; Jeng, 2005; Park, 2000
IF-consistency	Salampasis et al., 2002; Wesson, 2002; Zhang, 2004
IF-appropriateness (ADM, DEV, LIB, RES)	Zhang, 2004
IF-effort needed (USR, RES)	Jeng, 2005; Larsen, 2000; Zhang, 2004
IF-interaction support (ADM, DEV, LIB, USR)	Peng et al., 2004
UR-successfulness	Wildemuth et al., 2003; Zhang, 2004
UR-satisfaction	Bishop et al., 2000; Bollen & Luce, 2002; Cullen, 2001; Wilson & Landoni, 2001
UR-efficiency	Jones et al., 2002; Larsen, 2000; Meyyappan et al., 2004; Shim, 2000
UR-acceptance (ADM, DEV, LIB, RES)	Bollen & Luce, 2002; Mead & Gay, 1995
UR-productivity (RES, USR)[4]	
UR-use/reuse (ADM, DEV, LIB, URS)	Abbas et al., 2002; Baldwin, C., 1998; Bekele, 2002; Bishop, 1998; Bollen & Luce, 2002; Borghuis et al., 1996; Brophy et al., 2000; Carter & Janes, 2000; Cullen, 2001; Entlich et al., 1996; Hauptmann et al., 2001; Jones et al., 2000; Khalil & Jayatilleke, 2000; Lankes et al., 2003; Larsen, 2000; Marchionini, 2000; Monopoli et al., 2002; Shim, 2000
SV-accessibility	Lankes et al., 2003; Cullen, 2001
SV-integrity to search path	
SV-reliability	Cullen, 2001
SV-responsiveness	Cullen, 2001; Lankes et al., 2003; White, 2001
SV-usefulness	
CX-sustainability[5]	
CX-collaboration	
CX-managerial support	
CX-copyright compliance (ADM, DEV, LIB, RES)	Jones et al., 1999
CX-extended social impact (RES)	
CX-network effect (ADM, USR)	

[1]Proposed by Kwak et al., 2002; Kengeri, 1999

[2, 3] Proposed by Kwak et al., 2002

[4]Proposed by Lyman, 1997

[5]Proposed by Blixrud, 2002; Lynch, 2003

In addition to the context and technology level evaluation criteria, two content evaluation (*collection comprehensiveness* and *integrity*), one interface criterion (*supportiveness to HCI*), and two service level criteria (*integrity to information seeking path* and *usefulness to target users*) have not been examined in any DL evaluation studies.

Two issues are associated with the gaps. Firstly, there might be a lack of understanding of the importance of a given issue, such as *collaboration/sharing within/among different stakeholders*, *extended social effect*, is to a DL. Secondly, it might be difficult to develop a valid instrument to assess a given criterion. For instance, it might not be feasible to evaluate content *comprehensiveness* and *integrity* to other resources, because there seems to be no way to know how many documents can be considered as comprehensive in a given subject area, and what is out there that a given record/document/collection can be integrated with. Therefore, further efforts are needed to study these missing important criteria and carefully develop a valid methodology to assess them in real DL settings.

9.2 The Validity and Value of the Proposed Holistic DL Evaluation Model

The holistic DL evaluation model has been constructed based upon the series of research findings. Through the three dissertation stages, the author is able to identify 37 important criteria from the original pool of about 90 metrics and construct the model. Only very few criteria in the model are new to the existing body of research. The model should be able to serve as one of the most comprehensive models for DL evaluation for these reasons: (1) being in light of two promising conceptual DL evaluation frameworks, that is Marchionini (2000; 2003)'s multifaceted approach for inclusion of various groups of stakeholders in the research and Saracevic's (2000) stratified standpoint for

categorizing DL aspects; (2) relying on different and complementary research methods (i.e., interview, online survey, and experiment) for the three research stages of which the research purposes and instruments are interrelated and interdependent, and the results are complementary; (3) being grounded on the perspectives of various stakeholder groups (i.e., administrators, developers, librarians, researchers, and general users); and (4) being derived from the consistent results across the research stages.

Theoretically, through the process of the model construction, the research is able to contribute to the DL research body in these ways: (1) the research identifies more comprehensive and DL specific constructs and contexts. Within the original DL context level as pinpointed by Saracevic (2000), DL activity has been identified as an essential DL construct. In relation to the essential element, *collaboration/sharing* on DL development and product and *managerial support* have been highly regarded for DL evaluation at the context level; (2) the research further examines the divergence among various DL stakeholders in terms of what should be used for DL evaluations at different levels. The divergence, in particular, exists between the user and the other stakeholder groups. Meanwhile, the three library professional groups (i.e., administrators, librarians, and developers) tend to have more in common in their perceptions on what should be used for DL evaluation; (3) the research generates a comprehensive benchmarking framework for DL evaluations across systems, towards various directions and for different purposes. The research can likely fill gaps in current DL research area, especially where little has been addressed regarding what kinds of differences exist among various DL stakeholders in perceiving DLs, and how DL evaluation should be effectively conducted by soliciting diverse stakeholders' input and reflecting more DL

specific characteristics. In general, the three aspects of the contribution support earlier researchers' (Harter & Hert, 1997; Nicholson, 2004; Marchionini, 2000; Saracevic, 2000) arguments for multifaceted and multilevel evaluation.

Pragmatically, the proposed holistic DL evaluation model can provide DL developers and assessors with a comprehensive and flexible toolkit for conducting systematic DL design and evaluations. By using the toolkit, the DL professionals can readily conduct tailored DL evaluations for various purposes and with multiple perspectives. In other words, as suggested by the model, DL evaluation at a given level could be conducted by adopting all its core evaluation criteria and select some case-by-case criteria based upon evaluation objectives and target stakeholders' interests. For instance, when evaluating a DL interface, one should inclusively assess *ease of use, effectiveness,* and *consistency*. Meanwhile, he/she might also want to assess *effort needed* and *supportiveness to HCI* when aiming to include users' perspectives. Of course, he/she might also want to include *appropriateness to target users* for an internal evaluation.

In general, as pinpointed by Nicholson (2004), multiple criteria are needed to examine holistically the entire DL system, and individual criteria can be and should be integrated to produce a holistic view of a DL. This is the fundamental rationale for the core objective of this research. Additionally, as suggested by Harter and Hert (1997) and Marchionini (2000), a sound evaluation needs to have a good justification of evaluation criteria and balance of various stakeholders' interests. The proposed holistic DL evaluation model provides a feasible framework for such justification and balance.

9.3 Future Research

The generalizability of the model might be weakened by the academic setting of data sources. Some inconsistent results between the experiment and the other two stages also suggest the necessity for further verification. Meanwhile, the exclusion of funders' opinions and the heavy weight on stakeholders' subjective thoughts might have negative impacts on the comprehensiveness of the framework. Additionally, the model describes what criteria can be and should be used for various levels of DL evaluation. But it lacks a methodology for applying these criteria to actual evaluation studies.

Accordingly, further studies are needed to overcome the weaknesses and deficiency of this research, including (1) continuously testing the model in more DL systems and with greater stakeholder participation as well as using benchmarking search tasks; (2) enriching the model by the inclusion of more diverse stakeholders' (e.g., funder) opinions; (3) testing the model in various and presumably even beyond academic settings and with real evaluations using other than university library Web sites; and (4) developing a methodological framework for supporting the operationization of these criteria as well as empowering the flexibility of conducting various tailored evaluations in light of the holistic DL evaluation model. Especially, for some new criteria, such as *collaboration* and *managerial support*, it is necessary to address the measurement issue. Additionally, it would be helpful to develop exemplar evaluation instances to demonstrate how to select appropriate case-by-case criteria to achieve specific goals.

In conclusion, through a series of examination of diverse stakeholder groups' viewpoints about DL evaluations at various levels, this dissertation research produces a holistic model with specific criteria that can be tailored for multifaceted and multilevel

DL evaluation. As suggested by the model and the other findings, DL stakeholders share a great number of important criteria for DL evaluation, while having a small number of criteria with inter-group divergence. Although more verification work is needed to further test the validity of the model, considering the consistent results across the three stages, the research outcome should have foreseeable contributions to DL innovations in terms of revealing the inter-group divergence among stakeholders, as well as providing a holistic and flexible framework for DL evaluations.

References

Abbas, J. (2002). Middle school children's use of the ARTEMIS Digital Library. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, Oregon, July 13-17, 2002,* 98-105.

Adams A, & Blandford, A. (2001). Digital libraries in a clinical setting: friend or foe? *Research and advanced technology for digital libraries: Proceedings of 5th European conference, ECDL 2001*, Darmstadt, Germany, September 4-9, 2001, 214-224.

Arms, W.Y. (2000). Digital Libraries, Cambridge, MA, MIT Press.

Arms, W.Y. (1995). Key concepts in the architecture of the digital library. *D-Lib Magazine*, July 1995. Retrieved April 03, 2003, from http://www.dlib.org/dlib

Auerbach C. F. & Silverstein, L.B. (2003). Qualitative data: An introduction to coding and analysis, New York: New York University Press.

Baldonado, M.Q.W. (1999). A user-centered interface for information exploration in a heterogeneous digital library. *Journal of the American Society for Information Science,* 51(3): 297-310.

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

Bekele, S. (2002). The role and impact of the digital library on capacity building in the developing world – a case study of the OSSREA digital library. *International Information & Library Review*, 34(2): 129-37.

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. *Research and Advanced Technology for Digital Libraries: Proceedings of the 6th European Conference, ECDL'02,* September 16-18, 2002, Paris, France. 91-106.

Bertot, J.C., & McClure, C.R. (2003). Outcome assessment in the networked environment: research questions, issues, considerations, and moving forward. *Library Trends*, 51(4): 590-513.

Besek, J. M. (2003) Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment. Strategies and Tools for the Digital Library. Retrieved on July 12, 2004 from http://www.clir.org/pubs/reports/pub112/contents.html/.

Bishop, A.P. (1998). Measuring access, use, and success in digital libraries. *The Journal of Electronic Publishing,* v.3 (December 1998), Retrieved on September 25 from http://www.press.umich.edu/jep

Bishop, A.P. (1999). Making digital libraries go: comparing use across genres. *Proceedings of the Fourth ACM Conference on Digital Libraries,* 94-103.

Bishop, A.P., Neumann, L.J., Star, S.L., Merkel, C., Ignacia, E., & Sandusky, R.J. (2000). Digital libraries: situating use in changing information infrastructure. *Journal of the American Society for Information Science*, 51(4): 394-413.

Bishop, A.P., Van House, A.A., & Buttenfield, B.P. (2003). *Digital Library Use: Social Practice in Design and Evaluation*. Massachusettes, Cambridge: The MIT Press.

Blandford, A. & Buchanan, G. (2002). Workshop report: Usability of Digital Libraries @ JCDL'02, *ACM SIGIR Forum*, 36(2): 83-89.

Blixrud, J.C. (2002). Measures for electronic use: the ARL E-Metrics project. Retrieved on 4/12/04 http://www.lboro.ac.uk/departments/dis/lisu/Blixrud.pdf

Blixrud, J.C. (2003). Assessing library performance: New measures, methods, and models. 24th IATUL Conference, Retrieved on 4/12/04 at http://www.arl.org/stats/newmeas/newmeas.html

Bollen, J. & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine,* 8 (6). Retrieved on April 13, 2004 on http://www.dlib.org/dlib

Borghuis, M., Brinchman, H., & Fischer, A. et al. (1996). *Tulip: Final Report.* Elsevier Science Publication.

Borgman, C.L. (1999). What are digital libraries? Competing visions. *Information Processing and Management,* 35(3): 227-243.

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

Borgman, C.L. (2002a). Challenges in building digital libraries for the 21st century. *Digital libraries: people, knowledge, and technology: Proceedings of 5th International Conference on Asian Digital Libraries, ICADL 2002*, Singapore, December 11-14, 2002 1-13.

Borgman, C. (2002b). Fourth DELOS Workshop. Evaluation of digital libraries: testbeds, measurements, and metrics. Retrieved on 4/16/2003 from www.dli2.nsf.gov/internationalprojects/working_group_reports/evaluation.pdf

Borgman, C.L., Gilliland-Swetland, A.J. & Leazer, G.H. et al. (2000). Evaluating digital libraries for teaching and learning in undergraduate education: a case study of the Alexander Digital Earth ProtoType (ADEPT). *Library Trends*, 49(2): 228-250.

Borgman, C.L., Leazer, G.H., & Gilliland-Swetland, A.J. et al. (2001). Iterative design and evaluation of a geographic digital library for university students: a case study of the Alexander Digital Earth ProtoType (ADEPT). *Proceedings of the Fifth European Confernece on Research and Advanced Technology for Digital Libraries, Darmstadt, Germany, September 4-8, 2001*

Bosman, F.J.M., Bruza, P.D., & van de Weide, Th.P. et al. (1998). Documentation, cataloging, and query by navigation: A practical and sound approach. *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'98,* September 21-23, 1998, Heraklion, Crete, Greece. 459-478.

Brophy, P., Clarke, Z. & Brinkley, M. et al. (2000). EQUINOX: library performance measurement and equality management system—performance indicators for electronic library services. Retrieved on September 20, 2004 from http://equinox.dcu.ie/reports/pilist.html#pis

Browne, P. & Gurrin, C.(2001). Dublin City University Video Track experiments for TREC 2001, Retrieved April 03, 2003, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Budhu, M. & Coleman, A. (2002). The design and evaluation of interactivities in a digital library. D-Lib Magazine, 8 (11). Retrieved on April 13, 2004 from http://www.dlib.org/dlib

Bush, V. (1945). As we may think. Retrieved April 20, 2003, from http://www.theatlantic.com/unbound/flashbooks/compter/bushf.htm

Carter, D.S. & Janes, J. (2000). Unobtrusive data analysis of digital reference questions and service at the Internet Public Library: an exploratory study. *Library Trends,* 49(2): 251-265.

Champeny, L., Borgman, C.L., & Leazer, G.H. et al. (2004). Developing a digital learning environment: An evaluation of design and implementation processes. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*. 37-46.

Chowdhury, G.C. & Chowdhury, S. (1999). Digital library research: major issues and trends. *Journal of Documentation,* 55(4): 409-448.

Chowdhury, G G, Chowdhury, S. (2003). *Introduction to Digital Libraries*. London : Facet Pub.

Chowdhury, S., Hobbs, B., & Lorie, M. (2002). A framework for evaluating digital library services. D-Lib Magazine, 7/8. Retrieved on April 15, 2004 from http://www.dlib.org/dlib

Computer Science & Telecommunication Board, National Research Council (1998). Design and evaluation. A review of the state-of-the-art. *D-Lib Magazine*, July/Aug. Retrieved April 03, 2003, from http://www.dlib.org/dlib/july98/nrc/07nrc.html

Covey, D.T. (2002*). Usage and usability assessment: library practices and concerns*. Washington, D.C.: Digital Library Federation Council on Library and Information Resources. Retrieved on 4/12/2004 from http://www.clir.org/pubs/reports/pub105/contents.html

Cullen, R. (2001). Perspectives on user satisfaction surveys. *Library Trends*, 49(4): 662-686.

Dillon, A. (1999). Evaluating on time: a framework for the expert evaluation of digital interface usability. Retrieved on April 1, 2004 from http://www.ischool.utexas.edu/~adillon/publications/evaluating.html

Ding, W., Marchionini, G., & Soergel, D. (1999). Multimodel surrogates for video browsing. *Proceedings of Digital Libraries '99: the Fourth Annual ACM Conference on Digital Libraries.* Berkeley, CA, August 1999. 85-93.

Dorward, J., Reinke, D. & Recker, M. (2002). An evaluation model for a digital library service tool. *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Library,* 322-323

Elliott, M. (1995). Digital Library Design for Organizational Usability in the Courts. Retrieved on March 25, 2005 at http://www.lis.uiuc.edu/gslis/allerton/95/elliott.html

Entlich, R., Garson, L., Lesk, M., & Normore, L. et al. (1996). Testing a digital library: user response to the CORE project. *Library Hi Tech*, 14 (4): 99-118.

Fox, E.A., Hix, D., Lucy, T.N., Brueni, D.J, Wake, W.C. & Heath, L.S. (1993). Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science,* 44(8): 480-491.

Fox, E.A. & Urs, S.R. (2002). Digital libraries. In Blaise, C. ed. Annual Review of Information Science and Technology. Medford, NJ: Information Today, 36: 503-589.

Fuhr, N., Hansen, P., Mabe, M. & Micsik, A. (2001). Digital libraries: a generic classification and evaluation scheme. *Research and advanced technology for digital libraries: Proceedings of 5th European conference, ECDL 2001*, Darmstadt, Germany, September 4-9, 2001, 187-199.

Fuhr N., Klas, C.P., Schaefer, A. & Mutschke, P. (2002). DAFFODIL: an integrated desktop for supporting high-level search activities in federated digital libraries. *Research and advanced technology for digital libraries: Proceedings of 6th European conference, ECDL 2002*, Paris, France, September 16-18, 2002, 597-612.

Gladney, H.M. et al. (1994). Digital library: gross structure and requirements: report from a March 1994 workshop, retrieved on September 13, 2004 from http://www.csdl.tamu.edu/DL94/paper/fox.html

Goodrum, A.A. (2001). Multidimensional scaling of video surrogates. *Journal of American Society for Information Science*. 52(2): 174-182.

Greenberg, J., Bullard, K.A., & James, M.L. et al. (2002). Student comprehension of classification applications in a science education digital library. *Research and Advanced Technology for Digital Libraries: Proceedings of the 6th European Conference, ECDL'02,* September 16-18, 2002, Paris, France. 560-567.

Greenstein, D. (2000). Digital libraries and their challenges. *Library Trends,* 49 (2): 290-303.

Harter, S.P., and Hert, C.A. (1997). Evaluation of information retrieval systems: approaches, issues, and methods. In M.E. William, ed. *Annual Review of Information Science and Technology.* 32: pp.3-94. Medford, NJ: Information Today.

Hartland-Fox, B. & Dalton, P. (2002). eVALUEd: an evaluation model or e-library development. Ariadne (31) Retrieved on April 1, 2004 from http://www.ariadne.ac.uk/issue31/evalued/

Hauptmann, A. & Jin, R. (2001). Video retrieval with the Informedia Digital Video System. TREC'01, Retrieved April 03, 2003, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Hee, M., IK, Y.Y. & Kim, K.C. (1999). Unified video retrieval systems supporting similarity retrieval. *Proceedings of Tenth International Workshop on Database and Expert Systems Applications,* 884-888.

Hennig, N. & Web Advisory Group (2002). "Big Test" usability test, MIT Libraries, Cambridge, MA. Retrieved on December 25, 2004 at http://macfadden.mit.edu:9500/webgroup/usability2002/big-test/index.html

Hidaka, T., Abe, T. & Kokogawa, T. (2001). NetLibra: an advanced digital library system based on CORBA. *TREC'01,* Retrieved April 03, 2003, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Hill, L.L., Carver, L., & Larsgaard, M. et al. (2000). Alexander Digital Library: user evaluation studies and system design. *Journal of the American Society for Information Science,* 51(3): 246-259.

Hill, L.L., Dolin, R., et al. (1997). User evaluation: summary of the methodologies and results for the Alexander Digital Library, University of California at Santa Barbara. In C. Schwartz et. (Eds.) *Proceedings of the American Society for Information Science (ASIS) 97 Annual Meeting*, Washington DC, November 1997, Medford, NJ: Information Today, pp.225-243, 369.

Huxley, L. (2002). Renardus: following the Fox from project to service. *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'2002,* September 16-18, 2002, Paris, France. 218-229.

Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri: International Journal of Libraries and Information Services*, 55(2/3): 96-121.

Jones, G.J.F., & Lam-Adesina, A.M. (2002). An investigation of mixed-media information retrieval. R*esearch and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'2002,* September 16-18, 2002, Paris, France. 463-478.

Jones, M.L.W., Gay, G.K. & Rieger, R.H. (1999). Project soup: comparing evaluations of digital collection efforts. *D-Lib Magazine*, 5 (11) Retrieved on 3/12/2003 from http://www.dlib.org/

Jones, S, Cunningham, S.J., & Mcnab, R. et al. (2000). A transaction log analysis of a digital library. *International Journal of Digital Library,* 3:152-169.

Jones, S. & Paynter, G.W. (2002). Automatic extraction of document keyphrases for use in digital libraries: evaluation and application. *Journal of the American Society for Information Science* & Technology, 53 (8): 653- 677.

Kantor, P. & Saracevic, T. (1999). Quantitative studies of the value of research libraries: a foundation for the evaluation of digital libraries. *Proceedings of the American Society for Information Science,* 407-419

Kassim, A. R.C., & Kochtanek, T. R. (2003). Designing, implementing, and evaluating an educational digital library resource. *Online Information Review,* 27(3): 160-168.

Kengeri, R. et al. (1999). Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, and NDLTD. I*nternational Journal on Digital Libraries,* 2 (2/3): 157-169.

Kenney, A.R., Sharpe, L.H., & Berger, B. (1998). Illustrated book study: digital conversion requirements of printed illustration. *Research and Advanced Technology for*

Digital Libraries: Proceedings of the Second European Conference, ECDL'98, September 21-23, 1998, Heraklion, Crete, Greece. 279-293.

Khalil, M.A. & Jayatilleke, R. (2000). Digital libraries: their usage for the end user point of view. *Proceedings of the National Online Meeting,* May 16-18, 2000, New York, NY, 179-187.

Khoo, M. (2001). Ethnography, evaluation, and design as integrated strategies: a case study from WES. *Research and advanced technology for digital libraries: Proceedings of 5th European conference, ECDL 2001*, Darmstadt, Germany, September 4-9, 2001, 263-274.

Khoo, M., Devaul, H., & Sumner, T. (2002). Functional requirements for online tools to support community-led collections building. *Research and Advanced Technology for Digital Libraries: Proceedings of the 6th European Conference, ECDL'2002,* September 16-18, 2002, Paris, France. 190-203.

Kling, R. & Elliott, M. (1994). Digital library design for organizational usability. *ACM SIGOIS Bulletin*, 15(2): 59 - 70

Kwak, B.H., Jun, W. & Gruenwald, L. (2002). A study on the evaluation model for university libraries in digital environments. *Research and Advanced Technology for Digital Libraries. Proceedings of the 5th European conference, Rome, Italy, Sep. 16-18, 2002.* 204-217.

Lankes, R. D., Gross, M., & McClure, C. R. (2003). Cost, statistics, measures, and standards for digital reference services: a preliminary view. *Library Trends*, 51(3): 401-413.

Larsen, R. (2000) The DLib Test Suite and Metrics Working Group: Harvesting the Experience from the Digital Library Initiative. Retrieved on 4/13/04 from http://www.dlib.org/metrics/public/papers/The_Dlib_Test_Suite_and_Metrics.pdf

Lesk, M. (1997). Practical Digital Libraries: Books, Bytes, & Bucks. San Grancisco, California: Morgan Kaufmann Publisher.

Licklider, J.C.R. (1965). *Libraries of the Future*, Cambridge, MA MIT Press.

Lindlof, T.R. (1995). *Qualitative Communication Research Methods,* California: Sage Publication.

Lyman, P. (1997). Digital documents and the future of the academic community. *ARL Conference on Scholarly Communication and Technology*, Retrieved April 03, 2003, from http://www.arl.org/scomm/scat/lyman.html

Lynch, C. (2003). Colliding with the real world: Heresies and unexplored questions about audience, economics, and control of digital libraries. In Biship, A.P., Van House, A.A., & Buttenfield, B.P. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Massachusettes, Cambridge: The MIT Press.

Ma, Y.F. & Sheng, J. (2001). MSR-Asia at TREC-10 Video Track: shot boundary detection task. Retrieved April 03, 2003, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Marchall, C.C., & Ruotolo, C. (2002). Treading-in-the small: a study of reading on small form factor device. *Proceedings of the Joint IEEE and ACM Conference on Digital Libraries*, JCDL'02. Portland, Oregon, July 14-18, 2002.

Marchionini, G. (2000). Evaluation digital libraries: a longitudinal and multifaceted view. *Library Trends*, 49(2): 304-333.

Marchionini, G., Plaisant, C., & Komlodi, A. (2003). The people in digital libraries: multifaceted approaches to assessing needs and impact. In Ann P. Biship et al. (ed.) *Digital Library Use: Social Practice in Design and Evaluation*. Massachusettes, Cambridge: The MIT Press. pp.119-160.

Marchionini, G., Scaife, R., & Crane, G. (2000). Final evaluation report on the Perseus Project publication model (1997-2000). Final report to FIPSE. Retrieved on 4/15/2004 from http://ils.unc.edu/~march/perseus/final_report.pdf

McClure, Charles R.; Lankes, R. David; Gross, Melissa; Choltco-Devlin, Beverly (2002) Statistics, Measures, and Quality Standards for Assessing Digital Reference Library Services: Guidelines and Procedures. Retrieved on July 12, 2004 from http://quartz.syr.edu/quality/

McClure, C. & Fraser, B. (2002). Identifying and measuring library activities/services related to academic institutional outcomes. Retrieved on 4/12/04, http://www.arl.org/stats/newmeas/emetrics/phase3/ARL.Emetrics.Outcomes.Proposal.Final.Jan.8.02.pdf

Mead, J.P. & Gay, G. (1995). Concept mapping: An innovative approach to digital library design and evaluation. *ACM SIGOIS Bulletin,* 16(2): 10-14.

Meyyappan, N., Chowdhury, G.G., & Foo, S. (2000). A review of the status of 20 digital libraries. *Journal of Information Science*, 26(5): 337-355.

Meyyappan, N., Foo, S., & Chowdhury, G.G. (2004). Design and evaluation of a task-based digital library for the academic community. *Journal of Documentation*, 60(4): 449-475.

Nicholson, S. (2004). A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Journal of Documentation*, 60(2): 164-182.

O'Day, V.L. & Nardi, B.A. (2003). An ecological perspective on digital libraries. In Biship, A.P., Van House, A.A., & Buttenfield, B.P. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Massachusettes, Cambridge: The MIT Press.

Oppenheim, C. & Smithson, D. (1999). What is the hybrid library? *Journal of Information Science,* 25 (2): 97-112.

Papadakis, I., Andreou, I, & Chrissikopoulos, V. (2002). Interactive search results. *Research and Advanced Technology for Digital Libraries: Proceedings of the 6th European Conference, ECDL'02,* September 16-18, 2002, Rome, Italy. 448-462.

Park, S. (2000). Usability, user preference, effectiveness, and user behaviors when searching individual and integrated full-text databases: implication for digital libraries. *Journal of the American Society for Information Science,* 51(5): 456-468.

Peng L.K., Ramaiah, C.K., & Foo S. (2004). Heuristic-based user interface evaluation at Nanyang Technological University in Singapore. *Program-Electronic Library and Information System,* 38 (1): 42-59.

Prown, S. (1999). "Detecting 'broke': usability testing of library Web sites", Retrieved on December 12, 2004 at http://www.library.yale.edu/~prowns/nebic/nebictalk.html

Purcell, G.P., Rennels, G.D. & Shortliffe, E.H. (1997). Development and evaluation of a context-based document representation for searching the medical literature. *International Journal on Digital Libraries,* 1 (3): 288-296.

Rui, Y., Gupta, A. & Acero, A. (2000). Automatically extracting highlights for TV baseball programs. *Proceedings of the ACM Multimedia 2000*, 105-115.

Salampasis, Michail; Diamantaras, Konstantinos I. (2002). Experimental user-centered evaluation of an open hypermedia system and Web information seeking environments. *Journal of Digital Information*, 2 (4). Retrieved on July 12, 2004 from http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Salampasis/

Sanderson, M. & Crestani, F. (1998). Mixing and merging for spoken document retrieval. *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'98,* September 21-23, 1998, Heraklion, Crete, Greece. 397-407.

Saracevic, T. (1996). Modeling Interaction in information retrieval (IR) – A review and proposal. *Proceedings of the 59th Annual Meeting of American Society for Information Science,* 3-9.

Saracevic, T. (1997). The stratified model of information retrieval interaction. Extension and approaches. *Proceedings of the 60th Annual Meeting of the American Society for Information Science,* 313-327.

Saracevic, T. & Covi, L. (2000). Challenges for Digital Library Evaluation. Proceedings of the 63rd Annual Meeting of *American Society for Information Science*, 37:341-350.

Saracevic, T. (2000). Digital library evaluation: toward an evolution of concepts. *Library Trends,* 49(2): 350-369.

Seadle, M. (2003). Outcome-based evaluation. *Library Hi Tech*, 21 (1): 5 – 7.

Seadle, M. & Peters, T.A. (2000). Project ethnography: an anthropological approach to assessing digital library services. *Library Trends*, 49 (2): 370-385.

Sfakakis, M., & Kapidakis, S. (2002). User behavior tendencies on data collections in a digital library. *Research and Advanced Technology for Digital Libraries: Proceedings of the 6th European Conference, ECDL'2002,* September 16-18, 2002, Paris, France, 550-559.

Shim et al. (2001). ARL E-Metrics phrase II report. Retrieved on 4/16/03 from http://www.arl.org/stats/newmeas/emetrics/phasetwo.pdf

Shim, W. (2000). Measuring Services, Resources, Users, and Use in the Networked Environment. Retrieved on July 5, 2003 from http://www.arl.org/newsltr/210/emetrics.html

Shim, W. & Kantor, P.B. (1999). Evaluation of digital libraries: a DEA approach. ASIS'99; *Proceedings of the 62nd the American Society of Information Science (ASIS) Annual Meeting,* 36: 605-615.

Smeaton, A.F., Over, P., & Costello, C.J. et al. (2002). The TREC2001 Video Track: information retrieval on digital video information. *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'2002,* September 16-18, 2002, Paris, France. 266-275.

Smeaton, A.F. & Paul, O. (2001). The TREC-2001 Video Track report. Retrieved April 03, 2003, from http://trec.nist.gov/pubs/trec10/t10_proceedings.html

Spink, A., Wilson, T., Ellis, D. & Ford, N. (1998). Modeling users' successive searches in digital environment: a National Science Foundation/British Library funded study. *D-Lib Magazine*, April 1998. Retrieved April 03, 2003, from http://www.dlib.org/dlib/april98/04spink.html

Star, S.L., et al. (2003). Transparency beyond the individual level of scale: convergence between information artifacts and communities of practice. In Biship, A.P., Van House,

A.A., & Buttenfield, B.P. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Massachusettes, Cambridge: The MIT Press.

Stevens, J. P. (2002). Applied multivariate statistics for the social sciences (4th ed.) New Jersey: Lawrence Erlbaum Associates, Inc. pp.703.

Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory, Procedures and Techniques*. Newbury Park, Ca: Sage Publications.

Sumner, T., & Dawe, M. (2001). Looking at digital library usability from a reuse perspective. *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, USA, June 24-28, 2002,* 416-425.

Sumner, T., Khoo, M., & Recker, M. et al. (2003). Digital libraries in the classroom: Understanding educator perceptions of "quality" in digital libraries. *Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries*, 269-279.

Tabachnick, B.G. & Fidell, L.S. (2001). Using multivariate statistics (4th ed.) Needham Heights, MA: Allyn & Bacon. pp.966.

Thebridge, S., Dalton, P, & Hartland-Fox, R. et al. (2002). Outcomes assessment. Retrieved on 3/12/2004 from http://www.ebase.uce.ac.uk/evalued/Library/Paper2_Outcomes.pdf

Thong, J.Y.L., Hong, W.Y., & Tam, K.Y.(2002). Understanding user acceptance of digital libraries: What are the roles of interface characteristics, organizational context, and individual differences? *International Journal of Human-Computer Studies,* 57: 215-242.

Van House, N.A., Butler, M. and Schiff, L. (1996). Needs assessment and evaluation of a digital environmental library: the Berkeley experience. *DL96: the First ACM International Conference on Digital Libraries*, Bethesda, MD, March 20-23, 1996.

Van House, N.A., Butler, M.H., Ogle, V. & Schiff, L. (1996). User-centered iterative design for digital libraries: the Cypress experience. *D-Lib Magazine*, Feb 1996. Retrieved April 03, 2003, from http://www.dlib.org/dlib/february96/02vanhouse.html

Wallace, D.P. (2001). The nature of evaluation. In Danny P. Wallace & Connie Wan Fleet (ed.) *Library Evaluation: a Casebook and Can-Do Guide,* Englewood, CO: Libraries Unlimited, Inc. pp.209-220.

Waters, D. (1998). DLF Annual Report 1998-1999: Introduction. Retrieved on September 1, 2005 from http://www.diglib.org/ar9899p1.html

Wesson, J., & Greunen, D.V. (2002). Visualization of usability data: measuring task efficiency. *Proceedings of South African Institute of Computer Scientist and Information Technologists (SAICSIT) 2002*, 11-18

White, M.D. (2001). Digital reference services: framework for analysis and evaluation. *Library & Information Science Research*, 23 (3): 211-231.

Wildemuth, B., Marchionini, G., Yang. M., Geisler, G., Wilkens, T., Hughes, A., & Gruss, R. (2003). How fast is too fast? Evaluating fast forward surrogates for digital video. *Proceedings of the ACM/IEEE Joint Conference on Research on Digital Libraries* (Houston, TX: May 27-31, 2003), Los Alamitos, CA: IEEE. pp. 221-230.

Wilson, R. & Landoni, M. (2001). Evaluating electronic textbooks: a methodology. *Research and advanced technology for digital libraries: Proceedings of the 5th European conference,* Darmstadt, Germany, September 4-9, 2001, 1-12.

Xie, H.I., & Wolfram, D. (2002). State digital library usability: contributing organizational factors. *Journal of the American Society for Information Science & Technology*, 53(13): 1085-1097.

Xie, H.I. (2006). Evaluation of digital libraries: criteria and problems from users' perspectives. *Library and Information Science Research,* 28, 433-452.

Zhang, H.J., Low, C.Y., Smoliar, S.W. & Wu, J. H. (1995). Video parsing, retrieval and browsing: and integrated and content-based solution. *Proceedings of the third ACM international conference on Multimedia,* San Francisco, CA.

Zhang, Y. (2004) Moving Image Collections Evaluation – Final Report. Retrieved on October 25, 2004 from http://www.scils.rutgers.edu/~miceval

Zhang, Y., Jeng, J., & Li, Y.L. (2004). IFLA FRBR as User-Centered Metadata Evaluation Framework for Moving Image Collections. *Proceedings of Annual Meeting of American Society for Information Science & Technology*. Retrieved on November 1, 2004 from http://www.asis.org/Conferences/AM04/posters.html

APPENDICES

Appendix 1 Solicitation Letter to Candidate Interview participants (The Exploration Stage)

Dear Dr. (Mr. or Ms.)_____,

Will you please do me a favor?

For my dissertation research, I am conducting an interview with experts in the digital library research (development or administration). The purpose of the interview is to identify your perspectives regarding what criteria should be used for digital library evaluation. Your verbalized thoughts and opinions will help me in developing a generic model for digital library evaluation.

Considering your extensive expertise in digital library research (development or administration), your perspectives are important to the validity of the model.

The interview will take about one hour in whatever place you feel comfortable (i.e. your office, my office, or someplace else). During the interview, you will answer several questions pertaining to the research objective. Audiotaping will be used to record your verbalized thoughts and opinions.

This study has been approved by the Rutgers University Institutional Review Board. Surely, all your answers will be kept confidential. There will be no linkage to your name in the research product.

A copy of the research findings will be sent to you at your request.

Please simply reply this email with your preferred time and place for the interview. Thank you very much in advance for sincere consideration, time, and support.

Sincerely

Ying Zhang
Ph.D. Candidate
School of Communication, Information & Library Studies

Appendix-2 Consent Form for In-depth Interview (The Exploration Stage)

I understand that this interview aims to examine what criteria can and should be used in digital library evaluation from experts' point of view. Accordingly, I will talk about my personal opinions and perspectives. For the research purpose, the interview will be audio-taped and transcribed. I believe that the interview will be of little risk to me. Although I will not be paid for the interview, my opinions will benefit digital library research and development. I trust that the researcher will confidentially keep and use the interview data, and provide me with the final results before 12/31/2005 at my request. I also know that the interview will last about one hour, and I may withdraw from the interview whenever I would like to.

I, _____, consent to be interviewed on this day by the research, as well as grant the permit to the researcher to tape record and transcribe our conversation.

I would like to have the final results. ☐ Yes ☐ No

Signed,

Interviewee: _____ Date: _____

Interviewer: _____ Date: _____

Note: If you have any question or concern later on about the interview or the research, feel free to contact with either Ying Zhang (investigator) or Sponsored Program Administrator at the following addresses.

Ying Zhang
Ph.D. Candidate
Ph.D. Program of SCILS
Rutgers, the State University of New Jersey
4 Huntington Street
New Brunswick, NJ 08901
Phone: 732-729-0108
Fax: 732-932-2644
Email: yzhang@scils.rutgers.edu

Sponsored Programs Administrator
Office of Research and Sponsored Programs
Rutgers, the State University of New Jersey
ASB III, 3 Rutgers Plaza
New Brunswick, NJ 08901
Phone: (732) 932-0150 x.2104
Fax: (732) 932-0163
Email: szabo@orsp.rutgers.edu

Appendix-3 Interview Protocol (The Exploration Stage)

Introduction

Again, I appreciate your permission on the interview. The Institutional Review Board (IRB) at Rutgers wants me to make sure you experience no harm in any way. Please read and sign the consent form.

Experience/general knowledge about digital libraries

1. Digital library is a hot topic in library and information field. However, to date, there is no agree-upon definition regarding what is digital library. What comes to your mind if you are asked to provide a definition?

2. Could you please tell me about a couple of exemplar digital libraries that either you worked on (doing research/developing/manage) or you are familiar with?
   If work experiences → What are your experiences with the [***] digital library, in particular gains and lessons?
   If no work experiences → What are your impression about the [***] digital library, in particular its strengths and weaknesses?
   Probe: any more?

3. [Given the interview participant mentioned any evaluation experiences] Could you please tell me more about the [***] evaluation, for instance, what criteria were used and why they were selected for use?
   Probe: any more?

Personal perspective on DL evaluation criteria

> A scholar suggests that digital library can be evaluated at different levels, including content (information and collection), information technology (hardware and software), interface, service, user, as well as user's contexts (e.g. institutional, social, cultural). The following questions will ask you specifically criteria that should be used for each level.

1. **Content evaluation** examines how well digital collections are developed, digital objects are selected and created, and digital information/meta-information are organized and presented. If you were asked to evaluate digital content, including digital object, information, meta-information & collection, what criteria would you use?
   Probes: (1) could you explain it in detail? (2) any other criteria?

> Each of these cards shows a criterion that has been used and/or proposed by others for digital content evaluation. Could you please rank these cards by your perceived significance of each criterion to the content evaluation? You may refer to the back of a card for the notion of a criterion.

(a list of sorting cards with *criteria* and corresponding notions)

_____ *Accuracy* (the extent to which any visible errors (e.g. typo, incorrect information) are detected in digital information)

_____ *Appropriateness for target audience* (the extent to which the digital content is suitable for the domain knowledge and cognitive status of the target users)

_____ *Authority* (the extent to which the digital content is created by field experts or officials)

_____ *Conciseness of information* (the extent to which the digital information covers what need to be said in a short and clear manner without any unnecessary words)

_____ *Comprehensiveness of collection* (the extent to which the digital collection covers everything that is within a predetermined scope with respect to subject, time, language, format, etc.)

_____ *Ease of understanding* (the extent to which digital information can be easily understood)

_____ *Fidelity* (the extent to which digital information catches the detail and quality of originals)

_____ *Informativeness* (the extent to which the digital information about a topic has been well conveyed)

_____ *Scalability (of information)* (the extent to which digital information can be readily scaled up/down with diverse depth/ specificity so as to meet different user needs)

_____ *Timeliness (freshness)* (the extent to which digital information has been kept away from being out of date)

_____ *Usefulness to target user (*the extent to which digital information is helpful to target users to achieve their pre-determined goals*)*

2. **Technology evaluation** assesses how well hardware and software are developed and selected for supporting digital library searching. Given you were asked to evaluate digital technology for a given digital library, what criteria would you use?
Probes: (1) could you explain it in detail? (2) any other criteria?…

Please rank these cards by your perceived significance of each criterion to digital library technology evaluation? You may refer to the back of a card for the notion of a criterion.

(a list of sorting cards with *criteria* and corresponding notions)

_____ *Appropriateness for digital information* (the extent to which digital hardware is appropriate for digital information storage, process and display)

_____ *Comfort for use* (the extent to which a digital device is capable of being used in a manner that users feel comfortable)

_____ *Cost* (the amount of human and/or monetary resources needed to purchase/ develop digital hardware/software)

_____ *Display quality* (the extent to which digital information can be displayed technically at a high standard)

_____ *Effectiveness* (the extent to which software/hardware is devised to achieve pre-determined goals)
_____ *Efficiency* (the extent to which digital software/hardware can operate in a timesaving manner)
_____ *Flexibility* (the extent to which digital software/hardware is capable of being changeable in accordance with different situations)
_____ *Interoperability / compatibility among different systems* (the extent to which a given digital library can work together smoothly with other systems in a technical (software/hardware) sense)
_____ *Reliability* (the extent to which digital software / hardware can be trusted because they are capable of working without generating troubles)
_____ *Security* (the extent to which digital software/hardware is capable of protecting the system as well as user's personal information)

3. **Interface evaluation** mainly evaluates: (1) how effective and efficient a DL is in terms of helping users find information needed; (2) how well the interface fits users' knowledge background and information seeking needs/ behavior; and (3) how well the interface is in accordance with interface design principles. If you were asked to evaluate the interface of a digital library, including people's interaction with the interface, what criteria would you use?
Probes: (1) could you explain it in detail? (2) any other criteria?…

Please rank these cards by your perceived significance of each criterion to digital library interface evaluation? You may refer to the back of a card for the notion of a criterion.

(a list of sorting cards with criteria and corresponding notions)
_____ *Aesthetic attractiveness* (the extent to which a digital library interface is designed in a very pleasing manner aesthetically)
_____ *Appropriateness to target users* (the extent to which the interface is designed in a suitable way of meeting target users' background, needs and behavior)
_____ *Consistency* (the extent to which the interface is designed in a way that all essential elements (e.g. the color, layout, font, background, terminology use) are consistent across sections and pages)
_____ *Ease of use/learn* (the extent to which the interface is designed in a way that users can use (or learn to use) it easily)
_____ *Effectiveness* (the extent to which the interface is capable of helping users to achieve pre-determined goals and objectives)
_____ *Efficiency* (the extent to which the digital interface is designed in a timesaving manner)
_____ *Effort needed* (the amount of work required by the user to interact with the digital interface)
_____ *Error detection and handling capacity* (the extent to which the interface is capable of detecting and handling any errors that are caused by either users or the system)

_____ *Personalizability* (the extent to which the interface is designed in a way that a given target user can personalize the layout and content display in accordance with his/her preferences and needs.)
_____ *Supportiveness of H-C interaction* (the extent to which the interface is capable of providing assistance in human-computer interaction by visualizing interaction status)

4. **Service evaluation** measures how well a digital library may provide additional on-demand assistances (reference, tutorial, term suggestion, SSD-selective document dissemination) to users. If you were asked to evaluate digital service, what criteria would you use?
Probes: (1) could you explain it in detail? (2) any others?…

Please rank these cards by your perceived significance of each criterion to digital service evaluation? You may refer to the back of a card for the notion of a criterion.

(a list of sorting cards with *criteria* and corresponding notions)
_____ *Accessibility* (the extent to which users are free from barriers/restrictions (either physically or financially) to access a given digital service)
_____ *Cost-benefit* (the extent to which the digital service can furnish worthwhile positive outcomes against resource and time input after users utilize it)
_____ *Courtesy* (the extent to which the digital service is performed in a polite manner)
_____ *Empathy* (to extent to which the digital service is performed in a considerate manner that the staff understands users' feelings)
_____ *Gaps between expectation and perception* (what are the differences between users' expectations and actual perceptions?)
_____ *Number of positive feedback/reaction* (the amount of positive appraisals, appreciations, reuse intentions within a period of time)
_____ *Reliability* (the extent to which digital service can be trusted because of its minimized mistakes/errors)
_____ *Responsiveness* (the extent to which the service agent/staff is capable of providing users with positive and prompt responses)
_____ *Use/reuse* (the amount of usage, the number of returned users within a given period of time)

5. **User evaluation** primarily measures the outcome of digital library use in terms of any changes in human information behavior, cognitive, decision-making and or problem-solving capability, as well as any affective differences of a user or a group of users. Also, it measures impact/benefit on users' task in hand, and/or later on research, work, life, etc, due to the use of the digital library. If you were asked to do user evaluation, what criteria would you use?
Probes: (1) could you explain it in detail? (2) any other criteria?…

Please rank these cards by your perceived significance of each criterion to user evaluation? You mav refer to the back of a card for the notion of a criterion.

(a list of sorting cards with *criteria* and corresponding notions)

_____ *Acceptance* (the extent to which users express the willingness of using/reusing a given digital library)

_____ *Information literacy* (the extent to which users are able to improve their skills and analytic abilities to assess the validity and reliability of information sources after using a given digital library)

_____ *Learning effects* (the extent to which students, who have used a given digital library, are able to have their learning interests increased, critical thinking skills improved, and so forth)

_____ *Productivity changes before and after the digital library use* (the differences before and after the digital library use with respect to users' research/work outcomes)

_____ *Satisfaction* (the extent to which users have pleasant and satisfied feelings after using a digital library)

_____ *Successfulness of task completion* (the extent to which users complete their information search tasks successfully by using the digital library)

_____ *Time of task completion* (the amount of time users have spent on a given task while using the digital library)

_____ *Use/reuse* (the amount of usage (log-in sessions, document downloaded), and/or the amount of returned users)

6. Evaluation at the **context level** assesses how well a given digital library fits into larger contextual (e.g. institutional, social, cultural, economic, legal) practices, and what impacts and effects the digital library may have on these contextual practices. If you were asked evaluate DL's impacts at contexts, say social, cultural, legal, economical, what criteria would you use?
   Probes: (1) could you explain it in detail? (2) any other criteria?…

Please rank these cards by your perceived significance of each criterion to context evaluation? You may refer to the back of a card for the notion of a criterion.

(a list of sorting cards with *criteria* and corresponding notions)

_____ *Accessibility* (the extent to which community users are free of physical barrier, administrative, economic and social restrictions on using a given digital library)

_____ *Affordability/sustainability* (the extent to which it is affordable for an institute to develop/ subscribe/maintain a digital library)

_____ *Compatibility* (the extent to which a given digital library is capable of being integrated with other digital libraries and/or information retrieval systems within an institute in particular with respect to storing, editing and manipulating different files)

_____ *Copyright abidance* (the extent to which a given digital collection has no offense against copyright)

_____ *Integrity into organizational practices* (the extent to which a given digital library is capable of being smoothly integrated with organizational practices)

_____ *Number of incoming links* (the outreach impact in terms of the number of incoming links to a given digital library)
_____ *Outcome against predetermined institutional goals (*the extent to which the use of digital library yields to observable outcomes as being predefined in institutional goals)
_____ *Productivity of community members* (the extent to which the use of digital library yields to observable achievements (e.g. number of publications and/or grants) made by community members in an institute)

7. Any comments on the interview questions?

Appendix-4 Interview Transcripts Coding Rules (The Exploration Stage)

1. An analysis unit (quotation) should be the one that embraces a complete meaning. For some cases, it might be one sentence only, while for other cases, it might contain several sentences and even paragraphs (e.g. when an interviewee tries to further explain his/her thoughts).
2. Assign a quote to the code that has the closest meaning to it.
3. Assign a higher code to the DL level on which the interview question targets, unless there is an obvious meaning in the analysis unit for another DL level. For example, "To me, it not only accelerates the learning process..." should be assigned as "UR-Learning effects" rather than "CX-Extended social effects," although the interviewee was asked to express his/her thought about what criteria should be used for context evaluation.
4. When an analysis unit contains core words/phrases that are close or even exactly matching the code, just simply assign the code to it. Otherwise, figure out the main theme in the unit and assign the most closely matching code to it. For instance, "…the committee decided what would be the most **valuable** and **useful** across New Jersey to do as a top priority" should be directly coded as CT-Value and CT-Usefulness. However, "is the content robust enough--you know--I can find what I want, but there is nothing good in this database" could be coded as CT-Usefulness or CT-Value based upon a coder's interpretation.
5. If an interviewee denies the significance of a criterion, then no quotation as well as corresponding code(s) should be selected. For instance, "I don't think *easily understood* and *well-conveyed are critical"* should not be quoted and assigned with codes.
6. In general, use the smallest number of codes for an analysis unit based on the criteria that the codes should have the closest and/or the most specific meaning to the quote no matter whether or not this code belongs to the DL level the interview question addresses. Whenever possible, assign one code to a unit except for the rule 7 below.
7. In some cases, allow for a combination of two or more codes for a given quote if it is difficult to identify a single appropriate code for it. For instance, "And one might be I just found what I need, that is just very interesting as opposed he really looks into the **impact** you really want. I want to see what the **impact is**," can be coded as User-Productivity changes and UR-Learning effect, because the **impact** could have these two meanings.
8. For the units during the card-sorting stage, assign codes only to the top five criteria. For instance, "Ok, I'll rank [in order of importance] comfort of use, accessibility, display quality, appropriateness for digital information, inter-operationality, reliability, security, effectiveness, efficiency, and cost," only code *comfort of use, accessibility, display quality, appropriateness for digital information,* and *inter-operationality.*

Appendix-5 Interview Transcripts Coding Scheme (The Exploration Stage)

Code: CT-Accessibility
"the extent to which (full-text) information, meta-information is open for access"

Code: CT-Accuracy
"the extent to which any visible errors (e.g. typo, incorrect information) are detected in digital information"

Code: CT-Adequacy
"the extent to which a given DL may provide equal to or sufficient information for a specific requirement"

Code: CT-Appropriateness for target audience
"the extent to which the digital content is suitable for the domain knowledge and cognitive status of the target users"

Code: CT-Authority
"the extent to which the digital content is created by field experts or officials Associated with selectivity of content"

Code: CT-Comprehensiveness
“the extent to which the digital collection covers everything that is within a predetermined scope with respect to subject, time, language, format, etc."

Code: CT-Conciseness of Information
"the extent to which the digital information covers what need to be said in a short and clear manner without any unnecessary words"

Code: CT-Diversity
"The extent to which a DL has various formats of objects, such as full text, TOC, 3-D, manuscript..."

Code: CT-Ease of understanding
"the extent to which digital information can be easily understood"

Code: CT-Fidelity of information
"the extent to which digital information catches the physical and intellectual details of originals”

Code: CT-Informativeness
"the extent to which the digital information about a topic has been well conveyed"

Code: CT-Integrity of information over space/time
"The extent to which a DL information/collection can be incorporate with other resources over time and space"

Code: CT-Interest
"The extent to which information would be holding the attention of target users"

Code: CT-Ontological appropriateness
"the extent to which a DL and its components are mirrored to real worlds/ discipline represented without deviation."

Code: CT-Scalability of information
"the extent to which digital information can be readily scaled up/down with diverse depth/ specificity so as to meet different user needs”

Code: CT-Size
"the amount of documents, images, full-texts that are included in the DL"

Code: CT-Timeliness
"the extent to which digital information has been kept away from being out of date"

Code: CT-Uniqueness
"the extent to which digital information/collection are very unusually as appose to other existing DLs"

Code: CT-Usefulness to target users
"the extent to which digital information is helpful to target users to achieve their pre-determined goals "

Code: CT-Value
"The extent to which DL collection/information is worthy to target user/user group

Code: CX-Affordability/ Sustainability
“the extent to which it is affordable/able for an institute to develop/subscribe/maintain a digital library in financial, technical, political aspects. "

Code: CX-Collaboration/Sharing
"to what extent digital resources/technologies/services are sharing among DL stakeholders, including collaboration with other libraries, users, as well as collaboration among users”

Code: CX-Copyright abidance
"the extent to which a given digital collection has no offense against intellectual property protection"

Code: CX-Copyright reform/Fair Use
"The extent to which organizational accessibility is not limited by copyright laws"

Code: CX-Extended social impact

"The extent to which a DL might affect the practice of a given social group at an extended level that tends to be impossible before or unique in current practice, such as multi-disciplinary, the information poor, national, global, scholarly communication, preserving knowledge, technology dependency, education, democracy, legislation, economic"

Code: CX-Integrity into organizational/social practices
"the extent to which a given digital library (including its component) is capable of being smoothly integrated with organizational, social practices

Code: CX-Managerial support
"The extent to which DL development/maintenance is supported by human resource, physical resource, money resource, or vice versa...

Code: CX-Network effect
"The extent to which a DL is linked to other Web resources through incoming and/or outgoing links."

Code: CX-Organizational Accessibility
" the extent to which community users are free of physical barrier, administrative, economic and social restrictions on using a given digital library"

Code: CX-Outcome against predetermined institutional goals
"the extent to which the use of digital library yields to observable outcomes as being predefined in institutional goals
including fund raising"

Code: CX-Productivity of community members
"the extent to which the use of digital library yields to observable achievements (e.g. number of publications and/or grants) made by community members in an institute"

Code: CX-Scholarly communication
"The extent to which a DL is supportive to scholarly publication and research. "

Code: IF-Aesthetic attractiveness
"the extent to which a digital library interface is designed in a very pleasing manner aesthetically"

Code: IF-Appropriateness to target users
"the extent to which the interface is designed in a suitable way of meeting target users' background, needs and behavior"

Code: IF-Clarity
"To what extent the interface is designed in a way of being clear and free of cluttering"

Code: IF-Consistency

"the extent to which the interface is designed in a way that all essential elements (e.g. the color, layout, font, background, terminology use) are consistent across sections and pages"

Code: IF-Cost-Effectiveness
"the extent to which the interface can help in finding better results at lower (monetary, time) cost."

Code: IF-Distraction
"the extent to which a DL interface distracts the user himself/herself or the people next to him/her"

Code: IF-Ease of Navigation
"the extent to which the interface is designed in a way that users can easily get around from pages to pages, sections to sections"

Code: IF-Ease of use/learn
"the extent to which the interface is designed in a way that users can use (or learn to use) it easily, including intuitive, transparent, user-friendly"

Code: IF-Effectiveness
"the extent to which the interface is capable of helping users to achieve pre-determined goals and objectives"

Code: IF-Efficiency
" the extent to which the digital interface is designed in a timesaving manner as to general users"

Code: IF-Effort needed
"The extent to which a DL interface needs user's extra effort in order to interact with the digital interface to find desired information

Code: IF-Error detecting/handling capacity
" the extent to which the interface is capable of detecting and handling any errors that are caused by either users or the system"

Code: IF-Mimicry of real world
"The extent to which a DL can be mirroring real world.

Code: IF-Personalizability
"the extent to which the interface is designed in a way that a given target user can personalize the layout, function and content display in accordance with his/her preferences and needs."

Code: IF-Supportiveness of H-C/H-H/H-E interaction

"the extent to which the interface is capable of providing assistance in human-computer interaction by visualizing interaction status, facilitating users to take further action."

Code: SV-Accessibility
"the extent to which users are free from barriers/restrictions (either physically, financially) to access a given digital service. The restriction could be time, space, tool, copyright etc. "

Code: SV-Cost-benefit
"the extent to which the digital service can furnish worthwhile positive outcomes against resource and time input after users utilize it"

Code: SV-Courtesy
"the extent to which the digital service (staff) is performed in a polite manner"

Code: SV-Empathy
"to extent to which the digital service is performed in a considerate manner that the staff understands users¡¯ feelings"

Code: SV-Gaps between expectation & perception
"the extent to which users¡¯ expectation on what a digital service should be is different from (1) their perceptions on what service they are indeed provided or (2) the services that can be provided due to resource or administration restrictions.”

Code: SV-Integration to information seeking path
"To what extent a digital service is integrated into main flow of digital library use in a way of transparency so that it can be immediately reached by users.”

Code: SV-Positive feedback /reaction
"To what extent a digital service receives positive appraisals, appreciations, reuse intentions"

Code: SV-Reliability
"the extent to which digital service can be trusted because of its minimized mistakes/errors"

Code: SV-Responsiveness
" the extent to which the service agent/staff is capable of providing users with positive and prompt responses"

Code: SV-Use/reuse
" the extent to which a digital service is being used/reused by users, can be measured by the amount of usage, the number of returned users within a given period of time"

Code: SV-Usefulness to target users
"the extent to which a digital service is helpful to target users in achieving certain goals"

Code: TN-Appropriateness for digital information
"the extent to which digital hardware is appropriate for digital information storage, process and display"

Code: TN-Comfort for use
"the extent to which a digital device is capable of being used in a manner that users feel comfortable"

Code: TN-Cost
"The extent to which (human and/or monetary) resources needed to purchase/ develop/maintain digital hardware/software"

Code: TN-Display quality
"the extent to which digital information can be displayed technically at a high standard"

Code: TN-Ease of use
"the extent to which a digital technology is free of complexity so as to be used intuitively."

Code: TN-Effectiveness
"the extent to which software/hardware is devised to achieve desired effects"

Code: TN-Efficiency
"the extent to which digital software/hardware can operate in a timesaving/money-saving/effort-saving manner."

Code: TN-Flexibility
"the extent to which digital software/hardware is capable of being changeable in accordance with different situations

Code: TN-Interoperability / compatibility with other standards/systems
"the extent to which a given digital library can work together smoothly with other systems over time in a technical (software/hardware) sense, including federated search, with network facilities, with upcoming new technologies, ..."

Code: TN-Reliability
"the extent to which digital software / hardware can be trusted because they are capable of working without generating troubles

Code: TN-Robustness
"The extent to which digital technology is powerful enough to do most of things that users are looking for, such as interactivity, full-text search for plain text, video, audio, ..."

Code: TN-Security

"the extent to which digital software/hardware is capable of protecting the system as well as user's personal information (privacy)"

Code: TN-Standardization
"To what extent metadata are developed in a way of meeting updated standards

Code: TN-Timeliness
"To what extent a digital technology represents the mainstream at the time being, such as JAVA in the begin of the 21st century"

Code: UR-Absence of frustration
"The extent to which users are free of frustration while interacting with a DL"

Code: UR-Acceptance
"the extent to which users express the willingness of using/reusing a given digital library"

Code: UR-Behavior/workflow changes
"The extent to which user's information behavior (e.g. how to read) is changed due to the use of a DL "

Code: UR-Efficiency of task completion
"to what extent users' task on hand can be completed in a time saving manner"

Code: UR-Immersion
"to what extent users are plunged to a DL with increasing interests and tendency of use because they have very pleasing experiences with the library"

Code: UR-Information literacy
"the extent to which users are able to improve their skills and analytic abilities to assess the validity and reliability of information sources after using a given digital library"

Code: UR-Learning effect
"the extent to which students, who have used a given digital library, are able to have their grades improved, learning interests increased, critical thinking skills improved, knowledge gained in a more systematic manner, etc. through interactive learning process"

Code: UR-Value perception
"to what extent user(s) perceive the value of a DL"

Code: UR-Productivity changes before and after the digital library use
"the differences before and after the digital library use with respect to users¡¯ productivity in their life/work/research"

Code: UR-Satisfaction
"the extent to which users have pleasant and satisfied feelings after using a digital library"

Code: UR-Successfulness of task completion
"the extent to which users complete their information search tasks successfully by using the digital library

Code: UR-Supportiveness
"To what extent users show their understanding and support to a DL"

Code: UR-Use/reuse
"the extent to which a digital library and its components is used/reused by target/potential/ significant users (e.g. log-in sessions, number of documents downloaded, the amount of returned users, type of search queries, results in general, number of unusual users & other new uses...)"

Appendix-6 The Survey Questionnaire (The Confirmation Stage)

A digital library (DL) is an extension of a traditional library in terms of collection and function, as well as service. It provides users with various tools for finding and accessing different online resources, including books, journals, images, movies, sound recordings, etc. Additionally, it also provides virtual or human assistance (e.g. reference transactions) in case users need help in finding their desired information.

This survey was designed to gather your opinions to determine what criteria should or could be used to evaluate a digital library. The survey has seven sections. Sections A through F collect your perspectives on the significance of each criterion for six digital library aspects namely: content, technology, interface, service, user, and context. The final section asks essential demographic data primarily related to your searching experience. Your name and address will be collected only if you want to receive the special Thank-You gift for the first 100 survey participants ($3 value) and/or participate in drawing for a brand new Toshiba Satellite L25-S1216 laptop ($700 value).

The survey will take you about 20 to 30 minutes to complete. Your perspectives are very important to my dissertation research on developing a holistic framework for digital library evaluation. I want to thank you very much in advance for your support.

Section A--Content evaluation examines how well a digital collection is developed, how well digital objects such as images and sound recordings are selected and created, and how well digital information/meta-information such as full text articles and citations are organized and presented. Please click the most appropriate radio button to indicate each criterion's significance to the digital content evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text (e.g. Accessibility).** At the end of this section, you are encouraged to enter criteria for the digital content evaluation that you feel may have been missing from the survey.

Accessibility	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Accuracy of information	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Appropriateness for target audience	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Comprehensiveness of a collection	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Conciseness of information	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Ease of	O	O	O	O	O	O	O	O

understanding	Insignificant at all 1	Somehow insignificant 2	Slightly insignificant 3	Neutral 4	Slightly Significant 5	Somehow Significant 6	Extremely significant 7	Don't know 0
Fidelity of information	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Integrity of information over time/space	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Usefulness of information	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
: Other criterion 1: _____								
Other criterion 1: _____								

NEXT

Section B--Technology evaluation assesses how well digital hardware and software are developed and selected for supporting functions (e.g. searching, browsing, scanning) of a given digital library. Please click the most appropriate radio button to indicate each criterion's significance to the digital technology evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text** (e.g. Display quality). At the end of this section, you are encouraged to enter criteria for the digital technology evaluation that you feel may have been missing from the survey.

Appropriateness to digital information	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Display quality	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Ease of use	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Effectiveness	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Efficiency	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Flexibility	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Interoperability with other systems	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Reliability	O	O	O	O	O	O	O	O

	Insignificant at all 1	Somehow insignificant 2	Slightly insignificant 3	Neutral 4	Slightly Significant 5	Somehow Significant 6	Extremely significant 7	Don't know 0
Security	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
: Other criterion 1: _____								
Other criterion 1: _____								

PREVIOUS NEXT

Section C--Interface evaluation assesses how well a digital interface fits into users' background knowledge and information seeking needs, how well the interface is in helping users find information they need, and how well it complies to general interface design principles. Please click the most appropriate radio button to indicate each criterion's significance to the digital interface evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text** (e.g. Consistency). At the end of this section, you are encouraged to enter criteria for the digital interface evaluation that you feel may have been missing from this survey.

	Insignificant at all 1	Somehow insignificant 2	Slightly insignificant 3	Neutral 4	Slightly Significant 5	Somehow Significant 6	Extremely significant 7	Don't know 0
Aesthetic attractiveness	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Appropriateness to prospective users	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Consistency	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Ease of use/ learn to use	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Effectiveness	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Effort needed	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Personalizability	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Supportiveness to human - computer interaction	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
: Other criterion 1: _____								
Other criterion 1: _____								

PREVIOUS NEXT

Section D--Service evaluation assesses how well a digital library may provide additional on-demand assistance (reference, tutorial, document dissemination) to users. Please click the most appropriate radio button to indicate each criterion's significance to the digital service evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text** (e.g. Courtesy). At the end of this section, you are encouraged to enter criteria for the digital service evaluation that you feel may have been missing from this survey.

Accessibility	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Courtesy	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Gaps between expectation and perception	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Integrity to information seeking path	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Reliability	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Responsiveness	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Usefulness to target users	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0

:
Other criterion 1: _____

Other criterion 1: _____

PREVIOUS NEXT

Section E--User evaluation assesses the outcome of digital library use by examining changes on the user side, such as affective, cognitive, behavioral, problem-solving and decision-making capabilities. User evaluation also assesses the benefits that users have gained in completing current tasks or later used in research, work, and life. Please click the most appropriate radio button to indicate each criterion's significance to user evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text** (e.g. Acceptance). At the end of this section, you are encouraged to enter criteria for user evaluation that you feel may have been missing from this survey.

Acceptance	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Behavior change	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Information literacy	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Learning effect	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Productivity in user's work, research or life	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Satisfaction	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Successfulness of task completion	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Efficiency of task completion	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Use/reuse	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0

:

Other criterion 1: _____

Other criterion 1: _____

PREVIOUS NEXT

Section F--Context evaluation assesses how well a digital library fits into larger contextual (e.g. institutional, social, cultural, economic, legal) scale and also assesses what impact the digital library may have on these contextual components. Please click the most appropriate radio button to indicate each criterion's significance to the context evaluation in your perspective. **For reference, definitions are provided for each criterion when you move the mouse cursor over the text** (e.g. Copyright abidance). At the end of this section, you are encouraged to enter criteria for the context evaluation that you feel may have been missing from this survey.

Affordability / sustainability	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Collaboration /sharing	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Copyright abidance	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Extended social impact	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Integrity into organizational practice	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Managerial support	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Network effect	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Outcome as to pre-defined organizational goals	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0
Productivity of community members	O Insignificant at all 1	O Somehow insignificant 2	O Slightly insignificant 3	O Neutral 4	O Slightly Significant 5	O Somehow Significant 6	O Extremely significant 7	O Don't know 0

:
Other criterion 1: _____

Other criterion 1: _____

PREVIOUS NEXT

Section-G Please tell me something about yourself especially your experience with digital libraries and other online resources.

Age: O <20 O 20 ~ 29 O 30 ~ 39 O 40 ~49 O >=50

Gender: O Male O Female

Your highest education:
O High school O College O Graduate O Doctorate

Your most specialized field:
O Science, in particular _____
O Social science, in particular _____
O Humanity, in particular _____
O Other: _____

The role that most applied to you in relation to digital library:
- O **Administrator** who is the supervisor and decision-maker in digital library development. He/she does not have to be involved in detailed digital library development;
- O **Developer** who has at least engaged one digital library development project with specific roles in either designing or implementing the digital library;
- O **Researcher** who has at least written one published paper about digital library, and/or at least has taught one course related to digital library topics;
- O **Librarian** whose primary role is to help general users in finding books, articles, and other resources in a more effective and efficient way;
- O **General user** whose primary role is to utilize digital libraries in finding books, articles, and other resources for his/her research, study, work, or other purposes;
- O **Other:** _____

Please list several representative online sources that you have been developing and/or using for finding books, journal articles, and other materials:
1. _____
2. _____
3. _____
4. _____

Your mostly used online library catalog: _____
O Rutgers University Libraries online catalog
O Others: _____

Your general experience with developing and/or using these online sources has been for:
O < 1 year O 1 ~ 3 years O > 3 years O Other: _____

Your general frequency of developing and/or using these online sources is:
O Daily O Weekly O Monthly O Annually
O Others: _____

Thank you! You have finished the survey. Below is the essential information necessary to send you my special 'thank you' gift. The information is also required for the chance to win the Toshiba Satellite laptop. (Please note that for the laptop, you would be required to pay for the postage only if your address is in Hawaii, Alaska, or outside of the United States)

Yes, I would like to –

- □ Participate in the drawing for the Toshiba laptop;
- □ Receive Thank-You gift.

Here is my contact information:

Full Name: _____

Email for receiving notice: _____

Street and apartment number: _____

City, State, Zip code: _____

Country if outside of the United States: _____

PREVIOUS SUBMIT

Appendix-7 Experiment Consent Form (The Verification Stage)

You are invited to participate in evaluating Rutgers University Library (RUL) Web site at http://www.libraries.rutgers.edu, including IRIS (the library online catalog), the databases and electronic journals to which RUL has subscribed to, as well as other information/features on the Web site. The ultimate purpose of this research is to determine what criteria that should be used to evaluate the Web in your perspectives. Your participation will last approximately one hour. The procedures include:

1. completion of a pre-search questionnaire, providing brief background information about your age, education, experiences of digital library search, as well as writing down clearly a search task, for which you are going to search at the library Web site during the evaluation session;
2. completion of navigating and searching at the library Web site for the task that you have specified. You should pay special attention to those Web site features that either support or encumber your search. Also, you are encouraged to pay special attention to the aspects/ features you like or dislike most;
3. completion of a post-search questionnaire, providing your perspectives on the significant criteria that you would use in evaluating the Web site.

You will be assigned with a random subject number for the study. Your name will appear only on a list of subjects, and will not be linked to the subject number assigned to you. The evaluators will keep all data collected from you confidentially.

There are no foreseeable risks to participate in this study. Although your participation may not benefit you directly, the study can contribute to digital library innovation. For the one-hour session, you will receive a compensation of $10.00 in cash. The participation in this study is voluntary. You may withdraw at any time during the experiment and as such you will receive corresponding amount of compensation at the hourly rate.

If you have any questions about the study procedures, you may contact Ying Zhang (the Investigator) at (732) 322-2683. For questions about your rights as a research subject, you may contact the Sponsored Programs Administrator at Rutgers University at (732) 932-0150 ext. 2104. You will be given a copy of this consent form for your records.

Sign below if you agree to participate in this research:

Participant: _____ Date: _____

Investigator:_____ Date: _____

Appendix-8 Experiment Pre-Search Questionnaire (The Verification Stage)

User Number: _____ **Date:**_____

Age: O <20 O 20 ~ 29 O 30 ~ 39 □ 40 ~50 □ >50

Gender: O Male O Female

The role that most applied to you in relation to digital library:

□ Administrator, *who is the supervisor and decision-maker in digital library development. He/she does not have to be involved in detailed digital library development;*

□ Librarian, *whose primary role is to help general users in finding books, articles, and other resources in a more effective and efficient way;*

□ Developer, *who has at least engaged one digital library development project with specific roles in either designing or implementing the digital library;*

□ Researcher, *who has at least written one published paper about digital library, and/or at least has taught one course related to digital library topics;*

□ General user, *whose primary role is to utilize digital libraries in finding books, articles, and other resources for his/her research, study, work, or other purposes;*

O faculty member O doctoral student O master student
O undergraduate student O others:_____

Your most specialized field:

O Sciences, in particular: _____

O Social sciences, in particular: _____

O Art & Humanities, in particular: _____

O Others:_____

Your experience with searching the Rutgers University Library Web site:

O <1 yr O 1-3 yrs O 4-6 yrs O >6 yrs
O others:_____

Your frequency of searching the Rutgers University Library Web site:

O daily O weekly O monthly O yearly
O others:_____

Please articulate (PRINT) your search task for which you are expecting to get information from the Rutgers University Library Web site:

Appendix-9 Experiment Post-Search Questionnaire (The Verification Stage)

User Number: _____ Date: _____

To what extent were you able to finish the search task **with success**?			
O Not at all	O Very little	O Somewhat	O To a great extent

To what extent are you satisfied with your search **experience** with the **Rutgers University Library Web site?**				
O Dissatisfied	O Somewhat dissatisfied	O Neutral	O Somewhat Satisfied	O Satisfied

To what extent are you satisfied with your search **results?**				
O Dissatisfied	O Somewhat dissatisfied	O Neutral	O Somewhat Satisfied	O Satisfied

Please **PRINT** at least five significant features about **the Rutgers University Library Web site** that either assisted or hindered your task performance today at the **library Web site**:

Features that assisted your search:

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

Features that hindered your search:

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

The below are some statements about six digital library aspects, including content, technology, interface, service, context, and user. Please check off the ones that you think to be **the most appropriate** in relation to your experience with the Rutgers University Library Web site for the task.

Section A – Digital Content/Information/Collection

Digital content (e.g. journal articles, images) should be readily **accessible.**

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital content should be **accurate** without visible errors, including typos and incorrect information.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital content should be **appropriate** to the domain knowledge and cognitive status of prospective users.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital collection should be with a **relatively comprehensive** coverage within a predetermined scope.

Ox Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital information should be **nicely integrated** with other existing and new resources within a given digital library or even beyond the library to form a complete unit as opposed to unrelated pieces of information.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital content should be **useful** to prospective users in terms of helping them in achieving their pre-determined goals.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital content should be created, organized, and presented in a clear manner so that it can be **easily understood.**

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital information should be **concise** in terms of covering what needs to be said in a short and clear manner without any unnecessary words

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital information should be of good **fidelity** in terms of catching the physical and intellectual details of its originals (e.g. a digital copy or citation should authentically represent the original image or article)

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Section B--Digital Interface
Digital interface should be **supportive to human - computer interaction** by visualizing interaction status, such as the number of relevant documents, and search terms users may use for finding more relevant information) O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital interface should be **appropriate** to the background, needs and behavior of prospective users. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital interface should be designed in a very **pleasing and attractive** manner. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital interface should be designed in a way that its essential elements (e.g. color, layout, font, background, terminology use) are **consistent** across sections and pages. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital interface should be **effective** enough in helping users in finding information needed. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know

Digital interface should be designed in a way that users can **use (or learn to use) it easily and intuitively.**

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital interface should not require **extra effort** of users when the users are interacting with the interface for finding desired information.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Digital interface should provide users with various layouts and functional options so that they are able to **personalize** them based upon their own preferences and needs.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Section C – Digital Service
Additional digital service (e.g. online reference) should be **smoothly integrated into the main flow of digital library use** so that it can be reached by users whenever and wherever they need. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital service should be **reliable** with minimized mistakes/errors. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital service should be performed in a **courteous** manner. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
A digital service agent or staff member should be able to provide users with **positive and prompt responses.** O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital service should be **useful** to prospective users in helping them to achieve certain goals (e.g. finding a conference paper). O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know

Section D – Digital Library User
Users should be able to **complete their information search tasks successfully** through using a digital library. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Users should have **pleasant and satisfied feelings** after using a digital library. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
With a given digital library, users should be able to **finish their information seeking tasks in a timely fashion.** O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
A digital library and its components should be **used/reused** by its prospective as well as unexpected users. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
There should be a noticeable **productivity change** in users' work, research and/or life before and after their digital library use. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know

Digital library causes **changes in users' information behavior** (e.g. how to access or read full text papers).

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

There should be noticeable **acceptance** by a digital library's users, that is their willingness of using/reusing the digital library.

O Not applicable to my case
O Yes, it is important in my case and the importance level is:
O 1 O 2 O 3 (the most important)
O No, it is not so important in my case and the unimportance level is:
O 1 O 2 O 3 (the most unimportant)
O I don't know

Section E – Digital Technology/system
Digital system should be operated in an **efficient** manner (e.g. prompt response to search queries). O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital technology (e.g. plug-in software, such as Adobe Reader for content display) should be **easy to use** without any complexity. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital software/hardware is **flexible** so that it is readily changed in accordance with different situations and needs (e.g. developers are able to adjust management software when needed). O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
Digital system should be **reliable** without generating technical problems, such as "can't connection to server." O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 Ox 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
A digital library should be **smoothly integrated with other information retrieval systems** so that users can search across various databases simultaneously on a single interface. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know

A digital system should be capable of **protecting systems' as well as users' personal information**, such as what topic a user is looking for, what are the user's account and password.

O Not applicable to my case

O Yes, it is important in my case and the importance level is:

O 1 Ox 2 O 3 (the most important)

O No, it is not so important in my case and the unimportance level is:

O 1 O 2 O 3 (the most unimportant)

O I don't know

Section F – Context
A digital collection should comply with **intellectual property laws.** O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
A digital library should have lots of **incoming and outgoing links in relation to other Web resources.** O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
An institute should have adequate (financial, human, and/or technical) resources in **sustaining** a digital library. O Not applicable to my case Ox Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know
A digital library should be able to **affect the practice of a given social group at an extended level**, such as multi-disciplinary research, social/economic status improvement. O Not applicable to my case O Yes, it is important in my case and the importance level is: O 1 O 2 O 3 (the most important) O No, it is not so important in my case and the unimportance level is: O 1 O 2 O 3 (the most unimportant) O I don't know

CURRICULUM VITA

Ying Zhang

2007	Ph.D. in Communication, Information and Library Studies, Rutgers, The State University of New Jersey
2007	Zhang, X.M., Li, Y. L., Liu, J.J. & Zhang, Y. Effects of browse design in digital libraries on users' browsing experience. *Proceedings of Libraries in Digital Age 07 Annual Conference*, *May* 2007, Dubrovnik and Mljet, Croatia.
2007	Saracevic, T. & Zhang, Y. Criteria in evaluation of use and usability in digital libraries. *Proceedings of Libraries in Digital Age 07 Annual Conference*, *May* 2007, Dubrovnik and Mljet, Croatia.
2006-	Research Librarian for Asian Studies, University of California, Irvine
2006	Zhang, X.M., Li, Y. L. & Zhang, Y. Impact of interaction design for search features in digital libraries on user searching experience. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval,* August 2006, Seattle.
2004-2006	Team member, IEEExplore Evaluation project,
2004	Zhang, Y., Li, Y.L, Jeng, J. IFLA FRBR as user centered metadata evaluation framework for Moving Image Collections. *Proceedings of 2004 American Society for Information Science and Technology*, November 13-17, 2004, RI: Providence, http://www.asis.org/Conferences/AM04/posters.html
2004	Luo, C.R., Zhou, Z.N., & Zhang, Y. Building digitized collection: theory and practice. *Proceedings of The 7th International Conference of Asian Digital Libraries (ICADL)*, December 13 - 17， 2004, Shanghai, China, p.165-173.
2004	Saracevic, T. & Zhang, Y. Human information behavior and digital libraries: Tools for a survey. *Proceedings of Libraries in Digital Age* 2004, May 25-29, 2004, Dubrovnik and Island of Mljet, Croatia.
2003-2006	East Asian Studies Librarian at Rutgers University Libraries
2003-2006	Assistant to the Editor-in-Chief of Information Processing & Management, An International Journal
2003-2004	Evaluation team manage (research assistant), the Moving Image Collection Evaluation, a NSF-funded project
2003	Zhang, Y., & Smulewitz, G. An evaluation of computer-supported collaborative serial Management: A Case Study. *Information Technology and Libraries*, 22(3): 14~20.
2001-2004	Graduate assistant, Rutgers University Libraries
2001-2003	Resource manager of Disaster Recovery at NJServes.org
2000	He, Y.F., & Zhang, Y. The virtual library of Web resource. *Journal of Academic Libraries*, 18(5): 31~34 (Chinese)
2000	Zhang, Y., & He, Y.F. The future of Web search. *Contemporary Library and Information Technology,* no.3: 40~43 (Chinese)
2000	Zhang, Y. Current awareness tools on the Internet. *Academic Resources and Information Research*, no.1: 35~37 (Chinese)

2000	Zhang, Y. Bio-science resources on the Internet. *Biology Bulletin*, 35(1): 4~6. (Chinese)
1999	Zhang, Y. Comparative studies on three Internet search tools. *Library and Information Professionals,* no.10: 39~42 (Chinese)
1999	Zhang, Y. Comparative analysis of some genetic and gene databases and the optimized search tactics. *Academic Resources and Information Research*, no.1: 38~40 (Chinese)
1998-2001	Senior research librarian/Web master, Zhongshan University, China
1998-1999	Visiting scholar at Ohio University Libraries
1998	Zhang, Y. Several connection channels to DIALOG and the corresponding optimized strategies. *Contemporary Library and Information Technology,* no.3: 45~48. (Chinese)
1998	Zhang, Y. Patent resources on the Internet. *Academic Resources and Information Research*, no.1: 35~36 (Chinese)
1998	Liu, Y. & Zhang, Y. Environment science databases in the DIALOG system. In *Basic and Applied Research on Environmental Science----Paper Collection of Environmental Science Institute of Zhongshan University,* 1998, p244~248 (Chinese)
1997	Lei, Y., & Zhang, Y. New OCLC FirstSearch. *Academic Resources and Information Research,* no.3: 55~56 (Chinese)
1995	Master of Library Science, Zhongshan University, China
199-1998	Research librarian, Zhongshan University, China
1986-1989	Instructor, Department of Aquaculture, Shanghai Fishery College, China
1986	Bachelor of Science, Marine Biology, Qingdao Ocean University, China