Developing a Holistic Model for Digital Library Evaluation

Ying Zhang
University of California, Irvine, CA 92623. E-mail: yingz@uci.edu

This article reports the author's recent research in developing a holistic model for various levels of digital library (DL) evaluation in which perceived important criteria from heterogeneous stakeholder groups are organized and presented. To develop such a model, the author applied a three-stage research approach: exploration, confirmation, and verification. During the exploration stage, a literature review was conducted followed by an interview, along with a card sorting technique, to collect important criteria perceived by DL experts. Then the criteria identified were used for developing an online survey during the confirmation stage. Survey respondents (431 in total) from 22 countries rated the importance of the criteria. A holistic DL evaluation model was constructed using statistical techniques. Eventually, the verification stage was devised to test the reliability of the model in the context of searching and evaluating an operational DL. The proposed model fills two lacunae in the DL domain: (a) the lack of a comprehensive and flexible framework to guide and benchmark evaluations, and (b) the uncertainty about what divergence exists among heterogeneous DL stakeholders, including general users.

Background

The World Wide Web, along with advanced computation technologies, catalyses digital library (DL) research and practices. The past decade saw an exponential increase in the number of ongoing and completed DL projects. However, compared with the growing number of DL projects, the overall quality of DLs is insufficiently studied and reported (Chowdhury & Chowdhury, 2003; Goncalves, Moreira, Fox, & Watson, 2007; Isfandyari-Moghaddam & Bayat, 2008; Saracevic, 2000; Xie, 2006, 2008). "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). In addition to the quantity issue (i.e., not every DL project has been evaluated, and not every project with evaluation has its entire DL aspects covered), the quality of DL evaluation is problematic. Evaluation approaches and criteria vary among the existing studies. It is hardly possible to benchmark evaluation findings. Furthermore, the majority of the studies adopt traditional information retrieval (IR) and library evaluation approaches and criteria for examining common features (e.g., information accuracy, interface ease of use). Few metrics reflect unique DL characteristics, such as variety of digital format. And few address the effects of a DL at higher levels, including the extent to which a DL fits into or improves people's daily work/life (Bearman, 2007; Saracevic, 2000).

Having acknowledged the lacunae, a number of professionals and scholars 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 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 …." Similarly, Borgman (2002) 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).

This article reports a three-stage research of developing a holistic model for DL evaluation. It starts with a summary of general background, literature review, and research objectives followed by a detailed methodology descriptions. The Finding section reports major results with a focus on illustrating the proposed model along with summarizing important criteria perceptions among heterogeneous stakeholder groups for different levels of DL evaluations. Finally, the Discussion section suggests implications of the research to DL innovation and directions for future studies.

Previous Studies

The review of previous studies is focused on what criteria and framework have been used in the DL evaluations.

Evaluation Criteria

DL evaluation criteria and measures employed or proposed in existing literature can be essentially grouped at six levels, namely, content, technology, interface, service, user, and context, as suggested by Saracevic (2000).

Despite the significance of digital content evaluation (Xie, 2006), this body of research seems to be a weaker area.

Received September 8, 2008; revised May 27, 2009; accepted July 27, 2009

 DOI: 10.1002/asi.21220

Few studies report their DL evaluation at this level. Essentially, criteria are employed to assess four types of digital content: digital object, metadata, information, and collection. Among these four types, digital objects seem to be the unique type to DLs and, hence, be evaluated under DL-specific criteria, such as *fidelity* (Kenney, Sharpe, & Berger, 1998) and *suitability to original artifact* (Goodrum, 2001; Jones & Paynter, 2002). The remaining three types have been evaluated with conventional criteria, including *accuracy*, *clarity, cost, ease of understanding, informativeness, readability, timeliness,* and *usefulness*. Additionally, *scalability for user communities* (Kengeri, Seals, Reddy, Harley, & Fox, 1999; Kenney et al., 1998; Larsen, 2000) tackles a crucial issue in DL innovation, which involves more diverse user communities with various backgrounds and changing needs.

Digital technology evaluation has two foci: hardware and software. The latter uses primarily conventional relevance-based *effectiveness* measures, while several studies adapt them to fit into digital and hypermediated circumstances (Hee, Yoon, & Kim, 1999; Salampasis & Diamantaras, 2002). As for hardware evaluation, *display quality* and *robustness for digital information* are frequently used to evaluate electronic and communication devices. Meanwhile, *reliability, cost,* and *response time* are used for both hardware and software evaluations.

Interface is the most heavily evaluated DL level. Moreover, compared with the other five DL levels, interface evaluations tend to have more ready-to-use frameworks and criteria checklists, such as Wesson's (2002) multiple view, Nielson's (1993) five measures and 10 principles, Dillon's (1999) TIME framework, and Mead and Gay's (1995) evaluation tool. Nevertheless, only Nielsen's (1993) usability test attributes (*learnability, efficiency, memorability, errors,* and *satisfaction*) receive wide adoptions (e.g., Prown 1999; Peng, Ramaiah, & Foo, 2004).

Digital service evaluations examine how well a DL can provide additional on-demand (especially human or human-like) assistance to users. Lankes, Gross, & McClure, (2003) identified six criteria for evaluating digital reference, namely, *courtesy, accuracy, satisfaction, repeat users, awareness,* and *cost*. Some other criteria from traditional library face-to-face service evaluations (e.g., *accessibility*, *courtesy, empathy, reliability, difference before and after service intervention*, *gaps between expectation and perception)* can also be found in digital services. Additionally, a couple of criteria specifically fit digital reference transactions featuring time lag and invisibility in communication, such as *responsiveness* (Cullen, 2001; Lankes et al., 2003; White, 2001) and *user's control* (White, 2001).

Evaluations at the user level indirectly measure DLs through examining attributes of their users, such as changes of their information behaviors, benefits as to users' tasks in hand, or later on research, work, and life. So far, most evaluations at this level focus on the use/usage and benefits of individual searching and learning. Frequently used user-level criteria include *session time, accuracy of task completion*, *acceptance, use/intent to use,* and *satisfaction.*

In practice, DL evaluation at the context level is another weak area, regardless of its importance as pinpointed by several leading scholars (Bishop, 1999; Marchionini, 2000; Saracevic, 2000). To date, only very few evaluations have examined to some extent the contextual effects of DLs, including *copyright compliance* (Jones, Gay, & Rieger, 1999) and *preservation and spreading of culture* (Places et al., 2007). In addition, *sustainability* was proposed to measure the extent to which the augmentation of a DL could be secured without eventually losing its vitality (Blixrud, 2002; Lynch, 2003).

In sum, DL evaluations have been largely focused on interface and user levels. Content and context levels receive little attention. Moreover, most of the criteria used are merely borrowed from the domains of traditional library and information retrieval system. There lacks DL-specific evaluation measures for examining, for example, how well DL information and collections are integrated with each other, to what extent different DLs are compatible with each other, how well DLs support social/group interaction among heterogeneous users utilizing hypermedia information, and whether there are any changes in users' daily work and lives that are associated with DL applications.

Evaluation Frameworks

People have been working on developing frameworks/models for benchmarking evaluations. Among these studies, only a few provide criteria for multiple dimensions of DL evaluation: Kwak, Jun, & Gruenwald's (2002) evaluation model, Fuhr, Hansen, Mabe, & Miosik's (2001) DELOS evaluation scheme and U.S. DLI Metrics Working Group's quantitative performance measures (Larsen, 2000). Additionally, several large-scale programs have developed generic evaluation models for libraries in the digital age, including UK's eVALUEd (http://www.evalued.uce.ac.uk/), EU's EQUINOX (http://equinox.dcu.ie/), and ARL's New Measures Initiative (http://www.arl.org/stats/newmeas/index.html), LibQUAL^{+TM} (http://www.libqual.org/) protocol, and the newly developed DigiQUAL in the NSF/NSDL context (Kyrillidou & Giersch, 2005).

Other frameworks are primarily proposed for a single level of evaluation. For instance, Dillon's (1999) TIME framework, Mead and Gay's (1995) evaluation tool, and Wesson and Greunen's (2002) usability indicators are devised specifically for interface assessment. And White's (2001) descriptive model is used for analyzing and evaluating digital reference services.

Not only should attention be given to what evaluation frameworks are proposed, it is also vital to know how they are developed in order to see whether a given framework is valid and transferable to different settings. Among the handful of DL evaluation frameworks, the majority of them are constructed via consolidating experts' opinions, reviewing existing DL constructs, projects, and evaluation criteria, or relying on the researchers' own perspectives. The validity of these frameworks is weakened by either the

DOI: 10.1002/asi

exclusion of end users' opinions or the limitation of DL level coverage.

This research aims to develop a holistic DL evaluation model with a set of criteria covering core DL aspects and embracing perspectives from heterogeneous stakeholders, including DL end users. Two theoretical frameworks shed light on the research. One is Saracevic's (1996, 1997, 2000) stratified information retrieval (IR) model; and the other is Marchionini's (2000, 2003) multifaceted approach for assessing DL impacts.

The stratified model views an IR system, including a DL, as an entity containing components at different levels: content, technology, interface, user, service, and context. The system functions through interactions among the stratified levels. The model depicts essential components of a DL in a comprehensive but also flexible manner. In his conceptualization paper for DL evaluation, Saracevic (2000) describes the stratified layers as the "Contexts for Evaluation" (i.e., social, institutional, individual user, interface, system, and content). In other words, although the model was originally proposed for traditional IR systems, it should be still fitting to guide DL research. While the stratified model outlines what can be evaluated, Marchionini's multifaceted DL approach is a complementary framework, suggesting how quality data can be collected, analyzed, and reported. Having tackled the complexity of DL development with diverse people and activities, the multifaceted approach suggests DL evaluations to be conducted through taking different viewpoints, using different approaches and from different dimensions, then integrating data, and finally reaching a conclusion. Included are the stratified and multifaceted approaches that form enlightening guidelines for developing a holistic DL evaluation model in which diverse people's perspectives towards all kinds of levels.

Research Objectives

The main purpose of this research is to develop such a holistic DL evaluation model. The model should have the following two meanings in terms of being holistic: (a) cover all DL levels, including digital content, technology, interface, service, user, and context; and (b) bring in perspectives from as many diverse groups of stakeholders as possible. Three research objectives are as follows:

- 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.
- 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.
- To test the validity of the model when it is applied to extant DL use and evaluation.

Methodology

To develop the holistic DL evaluation model, I applied a hybrid research approach combining both qualitative and quantitative methods. Specifically, a three-stage research approach (see Figure 1)—*exploration, confirmation,* and *verification*—was devised to identify as many and as various as possible of criteria that could and should be used in DL evaluation, and eventually to construct a valid model with the inclusion of important criteria perceived by various stakeholders. These three stages are conceptually and methodologically interrelated.

During the *exploration* stage, a representative literature review and a semistructured interview were employed to examine what criteria could and should be used in DL evaluation. Then, the criteria identified from the exploration stage were embedded into an online questionnaire during the *confirmation* stage. More respondents from more heterogeneous DL stakeholder groups were asked to rate the importance of each criterion. The author constructed the holistic model by using descriptive and inference statistical techniques. Finally, in the *verification* stage, the validity of the model was tested through stakeholders' interaction with a real DL.

The selection of the research methods is carefully planned with a consideration of being appropriate to corresponding research objectives as well as maximizing the strengths of each method. For example, a semistructured interview has strength in eliciting a person's tacit thoughts, particularly when he or she pertains rich knowledge on the topic (Lindlof, 1995), and, thus, is appropriate for exploring as many as possible experts' perspectives on what criteria are important to DL evaluation, an area not so well explored. Meanwhile, an online survey is more suitable for statistically confirming the significance of these criteria through perspectives from a larger amount of and more diverse groups of DL stakeholders. Additionally, as a research method with both qualitative and quantitative nature, open-ended questions in the survey can be used to enrich the criteria set.

Literature Review–The Exploration Stage

I reviewed the literature using the following procedures:

1. Identified and selected related sources that are likely to cover DL evaluation literature.
2. Constructed search statements and composed search queries to retrieve DL evaluation literature.
3. Selected papers from retrieved sets that cover DL evaluation frameworks, methodologies, or criteria.
4. Summarized the frameworks, methodologies, and criteria from the papers selected.

Identification of sources. Various DL related sources were searched to identify criteria that have been used or proposed in existing research and development. 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. Additionally, DL project Web sites (e.g., *Digital Library Initiatives, ARL E-Metrics, EU EQUINOX, UK eVALUEd*) also served as core sources.

DOI: 10.1002/asi

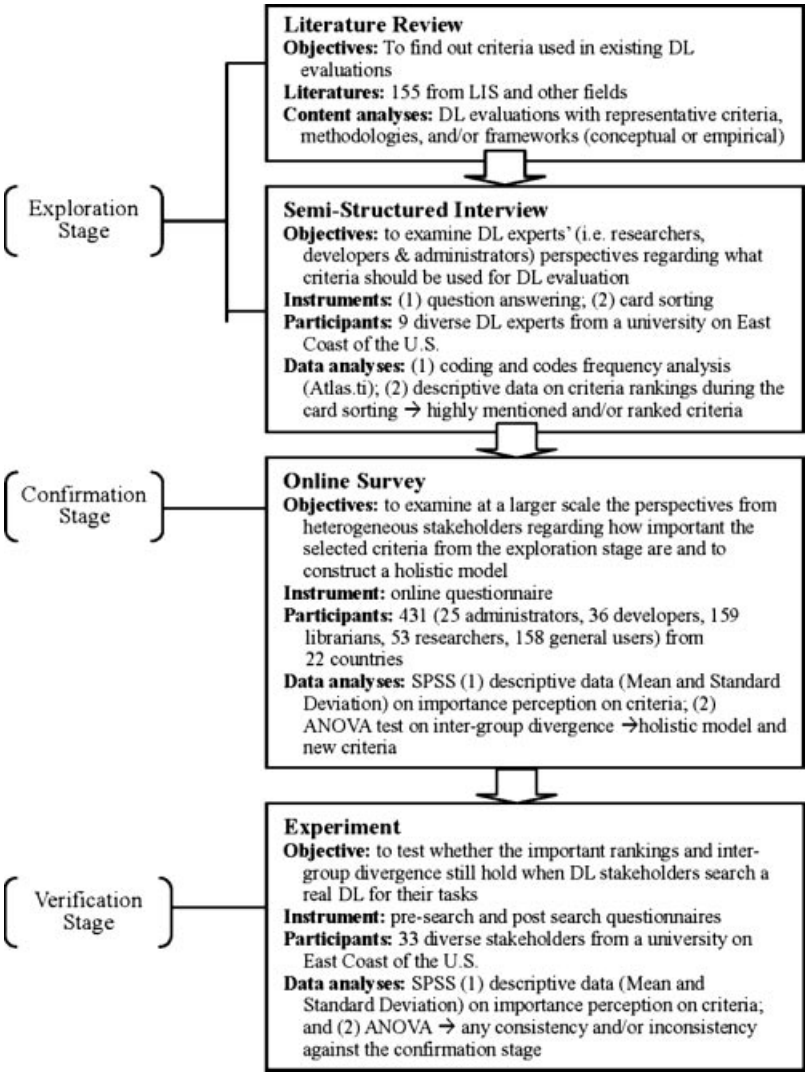

FIG. 1. Illustration of the three-stage research approach.

Considering the breadth of DL influences, Web of Science, a multidisciplinary database that indexes research articles from leading journals across disciplines, was also examined to expand the search scope to plausible DL application areas (e.g., education, health).

Search query composition. The primary search statement 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 depending on a given database's rules on search queries. Digital repository, as an emerging form for collecting, managing, and providing access to digital contents, was not used in the search query because it is too narrowly focused (Basefky, 2009) and has different boundaries than DL (Bearman, 2007). Additionally, it has been much less addressed in literature (a search in Web of Science on June 3, 2009 brings up 2,301 records for digital library/libraries with 1990 as the earliest publication year while merely 61 records for digital repository/repositories with 2001 as the earliest publication year). Meanwhile, the combination of "performance" and "outcome" with "evaluation" and "assessment" using the Boolean OR operator is simply for expansion of literature search scope.

Paper selection. The papers selected for the review were restricted to those studies with representative analyses or achievements on frameworks, methodologies, or criteria for DL evaluations. Eventually, 155 papers were selected as meeting the requirement. The justification for the selection is simply associated with the research objective, that is, to develop a holistic model for DL evaluation. Although the methodologies and frameworks primarily served as reference points with which this research was associated, the criteria identified from the literature were used in the later card sorting (CS) during the interviews, and for the discussion of new DL evaluation criteria identified from this research.

Criteria summarization. The literature review placed large effort on criteria identification, focusing on what criteria had been used for DL evaluation. The core results

DOI: 10.1002/asi

of the literature review—criteria lists that were used to develop an interview protocol in the succeeding research step—can be found at this persistent URL: http://hdl.rutgers.edu/1782.2/rucore10001600001.ETD.17104.

Semistructured Interview–The Exploration Stage

Interview participants. A purposive sampling method was employed to select nine DL experts who were likely to provide insightful thoughts about DL quality and performance indicators. Three groups of experts (i.e., administrators, developers, and researchers) with three in each group participated in the research. These expert stakeholders were recruited from the library school and the libraries at a university on the east coast of the United States. Interview participant eligibility required substantive knowledge of DLs and adequate experience developing, administering, or conducting research on DLs. Specifically, an eligible DL researcher should at least have published one paper or taught one course on DL. An eligible DL developer should have experience with designing or implementing at least one DL project. And an eligible DL administrator should be the one whose primary role is to oversee the implementation of at least one DL.

Having acknowledged the limitation of selecting the participants from a single institution, I made considerable efforts to increase the multiplicity of viewpoints by soliciting participants with varied backgrounds. For example, while one DL researcher participant was an expert on interaction in DL library, the other two specialized in technological and cultural aspects, respectively. Additionally, I employed two strategies to help the participants elaborate more DL evaluation criteria: (a) employing background-specific questions at the beginning as probes for getting them to think more about DL qualities later and (b) using the criteria identified from the literature in the CS as samples for eliciting more participants' own criteria.

Data collection. During June to October 2005, semistructured interviews were conducted to collect the nine DL stakeholders' perspectives on DL criteria. I interviewed each participant once for about an hour. At the beginning of each interview, he or she was asked to read and sign an interview consent form giving permission to be interviewed and to be audio-taped.

The nine interview questions were asked in the same order to minimize instrument bias (the instrument is available at http://hdl.rutgers.edu/1782.2/rucore10001600001.ETD.17104). After a couple of background-related questions were specific questions for eliciting participants' perspectives on the criteria that could, or should, be used in DL evaluation. Each of the DL criteria question targeted a given DL level, ranging from content, technology, interface, service, user to context.

For each DL level, in addition to the question-answering (QA) during which the interviewees spoke freely about DL evaluation criteria, a card-sorting (CS) technique was employed for them to rank several criteria that were preselected from the literature review results, based upon their occurrence frequency. The number of CS criteria for each level was restricted to 8-11 for manageable and meaningful results as inferred by an earlier pilot study. All participants sorted these cards based upon their perceived importance of each criterion to evaluation at the DL level. When sorting the cards, they were encouraged to refer to the back of a card for the definition of the criterion.

Data analysis. Qualitative data analysis software, Atlas.ti, was used to develop a coding scheme and to assign appropriate codes to meaningful narratives. The initial coding scheme was developed by incorporating results from the literature review and a pilot interview, and then it was applied in axial coding the nine interview transcripts. The scheme was organized into seven categories: one for DL constructs and six for the DL levels, as suggested by Saracevic in 2000. An open-ended coding technique was applied to identify new categories that were not included in the initial scheme. After the first coding run was finished, clean-ups were performed to remove less frequently mentioned criteria or to merge them to the closest ones if needed, in light of Auerbach and Silverstein's (2003) methodological suggestion.

To ensure coding consistency across transcripts, a set of rules was developed to guide the coding process. Additionally, the coder (i.e., myself) executed iterative coding-recoding reliability checking until two consecutive coding runs for each category reached a 70% or higher consistency rate. The coder repeated the coding runs independently (i.e., without referring to earlier coding results but with the same original coding scheme and with two consecutive coding runs apart from each other for at least 1 month). Additionally, the recoding processes were carried out only for those categories with less than 70% consistency rate against the previous run. Eventually, four coding runs were executed before all reached the reliability threshold.

Then, I 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 CS results was made to examine internal reliability within individual interviewees. I also sent the data analysis results back to the interviewees for "member checking" and received no requests for major changes. The criteria in the final list would be selectively included in the succeeding survey questionnaire for further confirmation by more respondents from more heterogeneous DL stakeholder groups.

Online Survey–The Confirmation Stage

Survey participants. Five groups of stakeholders participated in the online survey: researchers, developers, administrators, librarians, and general users. Whereas the general users were recruited from selected universities in the United States with LIS programs or active DL developments,

DOI: 10.1002/asi

TABLE 1. The listservs as sampling frames for the survey participants (confirmation stage).

Name	Description
ASIS_L	The listserv of American Society for Information Science and Technology
jSEES	The listserv of Association for Library and Information Science Education (ALISE)
ACRL_Forum	The listserv of Association of College & Research Libraries
LITA_L	The listserv of Library and Information Technology Association, a division of the American Library Association
LAMA_WOMAD	The listserv of women administrators from Library Administration and Management Association, a division of the American Library Association
LIBADMIN_L	Library administration discussion list, affiliated with the American Library Association
IFLA_L	The listserv of International Federation of Library Associations
IFLA_IT	The listserv of Information Technology Section, the International Federation of Library Associations
Web4Lib_L	An electronic discussion for library Web managers, hosted at University of California, Berkeley

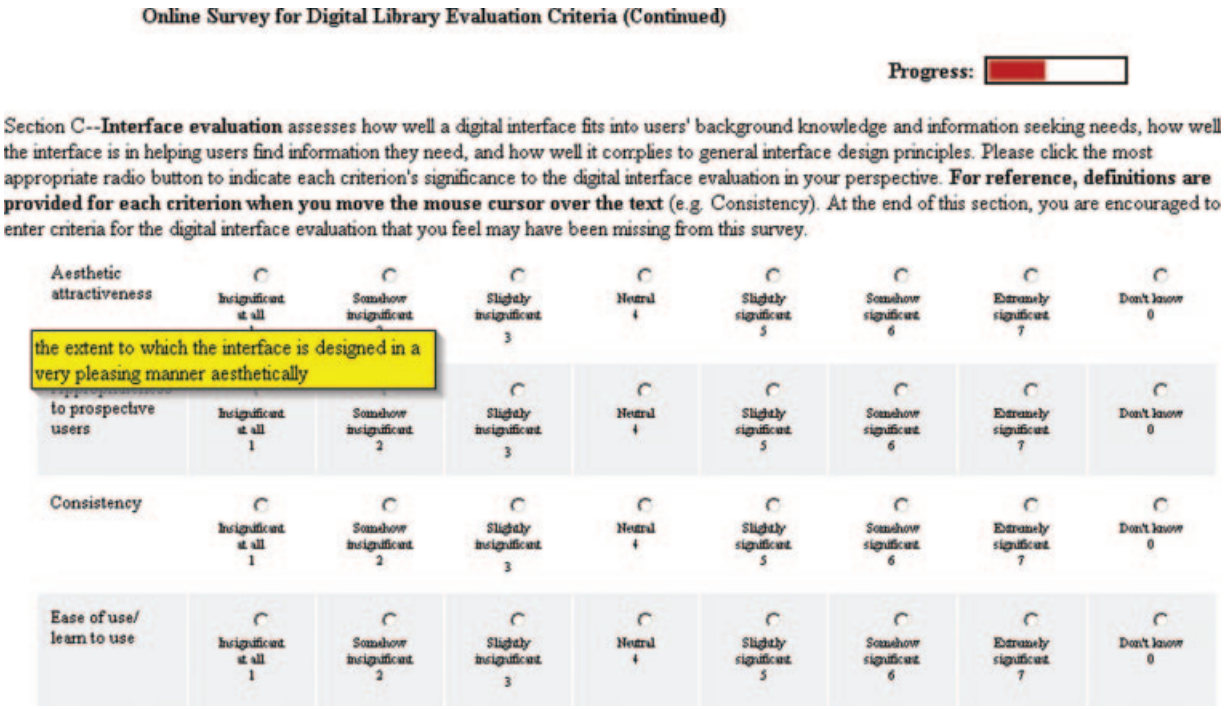

FIG. 2. Sample survey questions (confirmation stage).

the other four stakeholder groups were recruited from various academic and professional listservs. Table 1 provides a brief description about the listservs. The academic listservs tended to have more DL researcher members, and the professional ones might have more DL administrator, developer, and librarian participants. Meanwhile, the rationale for the sampling frame for general users was that faculty and students from those institutions tend to have more opportunity of using and becoming familiar with DLs and, thus, have more insightful 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 participants' self-reporting in the survey.

Data collection. During April to May 2006, an online survey recorded participants' perceptions on important DL evaluation criteria. A draw for digital devices and thank-you gifts were employed as incentives to increase response rate. A large percentage of survey respondents (87%) entered their names, as well as mail and e-mail addresses for receiving the gifts and result of the draw. The personal information suggests a low possibility of duplicate responses (i.e., very few people filled in the survey more than once).

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). Figure 2 demonstrates a sample survey section that shows the header and the first several questions. The header provided the explanation on the given level DL evaluation and the instruction on how to take the survey, including using a mouse-over action for the definition of a criterion (see the small box on the left of Figure 2 for an example) and entering additional criteria

DOI: 10.1002/asi

at the end of the section. The 7-point Likert scale ranged from 1 (*insignificant at all*) to 7 (*extremely significant*). "No opinion" option was also provided. In addition, alert pop-up windows were used to ensure that the participants finish all the sections and to prevent missing values. Progress bars indicated the finished/unfinished portion of the survey.

Data analysis. SPSS was used to analyze the data. Means and standard deviation were compared for a list of important criteria. Additionally, the one-way ANOVA test, a widely adopted statistical technique for examining differences in the mean values of a single variable associated with various groups, was conducted to examine inter-group divergence in perception of criteria importance (the single variable in this research context). ANOVA has strength in examining whether group means differ significantly and is weak in discovering which group means differ from one another. Therefore, wherever the inter-group divergence was identified, the post-hoc technique was employed to further identify the groups contributing to the divergence. The large sample size in the survey increases the robustness of the test departing from normality. The employment of the parametric test should not seriously violate the assumptions, as Glass, Peckham, & Sanders's (1972) finding suggests.

Experiment–The Verification Stage

Digital library system. The validity of the model constructed was tested through actual DL use. The Rutgers University Library (RUL) Web site (http://www.libraries.rugers.edu/) was the operational DL system used for testing. The choice was made because of the following two reasons: (a) the ease of getting experiment participants representing various and diverse stakeholder groups and (b) the likelihood of experiment participants' familiarity with the system as a benefit of being able to furnish more experience-based and knowledge-based perspectives on important criteria for DL evaluation.

As pointed out by several DL scholars (e.g., Borgman, 1999; Saracevic, 2000), thus far, there is no agreed-upon DL definition. The chaotic situation is also related to a debate regarding whether a library Web site can be considered a DL. In my viewpoint, a university Web site could be one type of DL for the following two reasons.

First, by comparing typical features on a representative library Web site (e.g., RUL) with the DL definition proposed by the Digital Library Federation (see Table 2), one may see that the former is essentially comparable with the latter. Specifically, the RUL site can be seen as the libraries' Web presence because the site contains a clear statement about the *organizational* mission, a well-defined *user community,* and a presentation of its *organizational structures* and resources. Meanwhile, it provides RU students, faculty, and other RU-*affiliated community members* with *readily and economically available resources*, including licensed databases and locally developed, rich digital collections (e.g., New Jersey Digital Highway) that are *selected*, *organized and integrated*, as well as *maintained* by *specialized staff*. From the site, faculty and students can not only search for physical library *collections* but also *access digital works.* Meanwhile, they may readily seek online *intellectual* assistance from *specialized librarians*.

TABLE 2. The digital library definition by the Digital Library Federation (Waters, 1998).

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.

Second, taking full advantage of network technologies, DLs gain their strengths by integrating distributed resources from different digital repositories. A DL does not necessarily require all its collections reside in a single local server. Furthermore, considering the enormous investments of a library on licensing commercial full-text e-resources and linking them to local systems, it is unfair to isolate all these resources from the Web site, which usually serves as the forefront of the library going digital. It is the integration of the Web site and the resources that makes a digital version of the library.

Experiment participants and their search tasks. During the summer of 2006, heterogeneous groups of stakeholders were recruited as experiment participants. The groups comprised general users, researchers, librarians, administrators, and developers. Whereas general users were recruited onsite in the two libraries (humanities and social science library and science and engineering library) in a university on the east coast of the United States, the latter four groups of participants were solicited through mailing lists of the university libraries and individual e-mail communications. As for the on-site recruitment, the author approached potential participants when they came to the library and started to use the library Web site. While the selection criteria for the groups of administrators, developers and researchers remained the same as the ones for the interview, the librarians were reference team members in the university libraries whose primary duty was to use the Web site to help and guide library users in finding information and library collections. These experiment participants were asked to prepare a search topic for locating relevant information via the library Web site.

Data collection. These participants' perceptions about important criteria for the library Web site evaluation were collected through a post-search questionnaire after they finished searching the site. The questionnaire included all criteria from the holistic DL evaluation model, plus a few that were perceived to be least important by the survey participants. The inclusion of the least important criteria serves as an additional examination on whether these criteria are still considered to be the least important in the real DL-use setting. 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 library Web site. A sample statement was: "Digital interface should be

DOI: 10.1002/asi

designed in a way that its essential elements (e.g., color, layout, font, background, terminology use) are *consistent* across sections and pages." The participants could select any of the three options, that is, "not applicable to my case," "I don't know", and importance rating on a 6-point Likert scale from 1 (*least important*) to 6 (*most important*).

In addition to the perceived importance rating on the preselected criteria, the participants were also encouraged to enter in open-ended sections for perceived significant features on the site that had helped (or hindered) them in the task implementation. They were encouraged to pay special attention, while searching, to information, interfaces, and various functions on the site rather than those provided by off-site commercial databases licensed by the libraries.

Data analysis. Again, SPSS was used to analyze the distribution patterns of the participants' importance ratings and to identify group differences among the stakeholders. The participants' "I don't know" answers were treated as missing data because they had no meaning on the importance of a criterion. Their "not applicable to my case" answers were coded as zero, and they were included in the frequency analysis, but not in the descriptive and inference analyses (ANOVA), because the inclusion could bias the mean score and enlarge the standard deviation (SD). Besides, these answers had totally different meanings from the importance ratings.

The results were compared with the ones from the confirmation stage for examining whether the important criteria from the confirmation stage were still perceived to be important when DL stakeholders interact with the operational DL and whether the inter-group differences still held.

Findings

The Findings section reports demographic data of the interview, survey and experiments participants, the similarity and divergence of their perceived most and least important DL evaluation criteria, and the proposed holistic DL evaluation model in terms of the similarity and divergence.

Research Participants

The interview participants—exploration stage. The author interviewed nine DL stakeholders, including three DL administrators (IA1, IA2, and IA3) and three DL developers (ID1, ID2, and ID3) from the libraries of a university on the east coast of the United States and three DL researchers (IR1, IR2, and IR3) from the Library and Information Science Program at the same institution.

The survey participants—confirmation stage. In total, 434 survey participants finished the survey, of which the data of 431 were usable. Of these 431 participants, 159 (37%) self-reported their primary roles as librarians and 158 (37%) 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. The difference in the group sample size is probably associated with population variance. Usually, the numbers of DL administrators, developers, and researchers are smaller than those of librarians and users.

About half of the survey participants (220, 51%) were 30 to 49 years old, 93 (22%) were over 50, and 118 (27%) were 20 to 29 years old. The gender distribution was 167 (38.7 %) male and 264 (61.3%) female and was almost equally distributed among the four stakeholder groups except for the librarians. The librarian group had more females (114, 72%) than males (45, 28%). In terms of the highest education level achieved, the majority of participants held graduate (308, 71%) or doctoral degrees (100, 23%). Only 23 (5%) had baccalaureate or lower degrees. The skewed education level might be associated with the sampling frames of university settings for users and of the academic and professional listservs for the other four groups who are more likely to hold higher degrees. The subject backgrounds showed 209 (48%) for social sciences, 130 (30%) for sciences, 79 (18%) for humanities and arts, and 13 (3%) for others. Most survey participants (314, 73%) had been searching online for more than 3 years.

The survey also attracted overseas participants. Among 367 (85%) participants who reported their nations, 310 (85%) were from the United States, and 57 (15%) were from 21 other countries, including China (16), United Kingdom (7), Germany (3), Greece (3), Spain (3), New Zealand (3), India (2), Egypt (1), Finland (1), Italy (1), Japan (1), Kenya (1), Korea (1), Mexico (1), Sweden (1).

The experiment participants and their search tasks—verification stage. Thirty-three DL stakeholders from a university on the east coast of the United States 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.

In terms of age distribution, more than half of the participants (19, 58%) 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 composition of subject fields for these participants was 11 (33%) for social sciences, 10 (30%) for sciences, and 7 (21%) for humanities and arts. More than half of the participants (19, 58%) had been using the university library Web for more than 3 years and over three-fourths (25, 76%) used the Web on a daily to weekly basis.

Although the participants came up with their own search tasks of various topics, their search tasks were essentially to either find books/articles/images/other Web resources about a given topic (26 cases, 79%) or locate known items (7 cases, 21%). There was little inter-group difference among them in terms of the types of search tasks.

The Most and Least Important DL Evaluation Criteria

The interview participants' perspectivess Table 3 lists the top five important and the three least important DL evaluation

DOI: 10.1002/asi

TABLE 3. Interview participants' top important and non-important evaluation criteria.*

	Important criteria		Non-important criteria	
	QA	CS	QA	CS
Content	**Usefulness** (32; 9) Accessibility (32; 7) Integrity (24; 6) Comprehensiveness(22; 6) **Ease of understanding** (20; 7)	**Usefulness** (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** (41; 9) Personalization (20; 7) **Effectiveness** (20; 6) **Appropriateness** (16; 9) Support of HCI (15; 6)	**Ease of use** (1.8) **Appropriateness** (2.3) **Effectiveness** (3.7) Consistency (5.3) Effort needed (5.6)	Free of distraction (3; 2) Mimicry of reality (7; 2) Attractiveness (8; 6)	Personalization (8.8) Support of HCI (7.3) Attractiveness (7.2)
Service	Integrity (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 compliance (5.1)	Network effect (6; 3) Outcome (6; 4) Productivity (9; 6)	Network effect (6.6) Compatibility (5.8) Organizational accessibility (5.2)

QA = question answering; CS = card sorting.
*The texts in bold are the important criteria that appeared in both the QA and the CS top-five rankings.

criteria from the open-ended QA and the CS, which is 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 interviewees in CS. The data are grouped into the six DL levels.

The criteria displayed on sorting cards for each DL level were preselected from the literature review findings. The number of CS criteria for each level was limited to 8-11. In contrast, there was no such preselection and restriction for QA criteria. What criteria and how frequently they were mentioned were 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 and collection, what criteria would you use?"

Furthermore, for a given DL level, CS always followed the open QA. Accordingly, criteria that were heavily mentioned by an interviewee might not be on the sorting cards. Similarly, during the open QA, an interviewee might not even mention a criterion highly ranked by him or her during CS. The transcripts revealed that some important criteria were excluded in the open QA 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…"

In addition to the recall effect, the variation in total number of criteria between CS and QA and the emergence 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) might be worth examining. Meanwhile, considering the primary research objective, which is to identify what criteria should be used for DL evaluation, the analyses focused more on the important criteria perceived rather than the unimportant ones. In general, over half of the important criteria (see the texts in bold in Table 3)—16 out of 30—appeared in both the QA and the CS top five rankings.

DOI: 10.1002/asi

TABLE 4. Survey participants' top five and least important criteria ($n = 431$).

	Content	Technology	Interface	Service	User	Context
Most	Accessibility 6.52 (1.00)	Reliability 6.49 (0.93)	Effectiveness 6.35 (0.99)	Reliability 6.39 (1.00)	Success 6.38 (0.98)	Sustainability 6.32 (1.05)
	Accuracy 6.53 (1.07)	Ease of use 6.35 (1.02)	Ease of use 6.33 (1.02)	Accessibility 6.29 (1.09)	efficiency 6.06 (1.07)	Collaboration 5.92 (1.10)
	usefulness 6.09 (1.19)	Effectiveness 6.21 (1.00)	Consistency 5.88 (1.16)	Usefulness 6.28 (1.06)	Satisfaction 6.07 (1.19)	copyright compliance 5.82 (1.58)
	Fidelity 6.04 (1.21)	Interoperability 6.05 (1.23)	Effort needed 5.88 (1.19)	Responsiveness 6.17 (1.08)	use/reuse 6.02 (1.13)	Managerial support 5.76(1.23)
	Integrity 5.97(1.17)	efficiency 6.03 (1.07)	Appropriate 5.83 (1.15)	Integrity 5.93 (1.17)	Productivity 5.94 (1.27)	Network effect 5.66 (1.29)
Least	Conciseness 5.14 (1.38)	Flexibility 5.64 (1.45)	Personalization 4.75 (1.46)	Courtesy 5.28(1.39)	Behavior change 5.13 (1.38)	Extended social impact 5.19 (1.41)

The survey participants' perspectives. Table 4 summarizes the five most important criteria and the lowest regarded criterion of each of the six DL levels perceived by the survey participants. The rankings of the importance rating are based upon descriptive data, which include the mean (outside the parentheses) as the primary factor and SD (in the parentheses) as the secondary. Only when two criteria are identical in mean scores, the role of SD comes to play. The larger is the mean and the smaller the SD, the higher the ranking.

Essentially, the important criteria perceived by the interviewees are also perceived to be significant by the survey participants, and so are the least significant criteria. For content level evaluation, *usefulness to target users* was consistently top ranked. It appeared in the top five lists of the survey, as well as the interview CS and QA. Unanimously, the interviewees and the survey participants regarded *conciseness of information* as the least important criterion. Digital technology evaluation criteria also were ranked consistently across the two studies. *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 QA section.

Similarly, for interface level evaluation, 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. *Attractiveness,* the least important criterion in the interview, was still the second lowest ranked in the survey. The results for service level evaluation are also consistent. In particular, *service accessibility* and *integrity to information-seeking path* appeared in the top five lists of the interview CS and QA as well as of the survey. Similarly, *courtesy* was the lowest-ranked criterion in the three lists, presumably because it does not directly influence users' search outcome.

In contrast to the high consistency in perceived important criteria at these four lower levels, DL evaluation criteria at user and context levels show a large variance between the interview and the survey results. For user level evaluations, while *successfulness*, *efficiency of task completion*, and *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*) were not at the top of the survey list. *Behavior changes* dropped to the lowest-ranked criterion. This is presumably associated with the inclusion of user groups in the survey. Users tended to care more about the direct effects of using a DL, such as *efficiency* and *successfulness of task completion*, and less about the indirect outcomes.

As for context evaluation, although the interviewees and the survey participants agreed on *sustainability* as the most important criterion for assessing a DL at its context level, they were unlikely to have parallel perception about the importance of DL's *extended social impact*. This criterion was highly regarded in the interview QA; but it became the least important criterion in the survey. Another highly ranked criterion (i.e., *integrity to social practice*) in the interview also dropped to the least second. In contrast, incoming and outgoing hyperlinks (i.e., *network effect*) in the survey participants' perspectives were important to a certain extent, whereas it was the lowest-ranked criterion in the interview QA and CS.

For the lower level DL evaluation, several instances of inconsistency between the two studies have also been observed. For example, technological *flexibility* was highly ranked in the interview CS, but in the survey it was the lowest-ranked criterion, possibly because the criterion seems to be of greater interest to DL developers than to users, and the inclusion of users' opinions in the survey contributed to the ranking drop. Additionally, the two lowest-ranked criteria in the interview (i.e., d*isplay quality* and *security)* were ranked more highly in the survey. For service level DL evaluation, there was only one inconsistent perception (i.e., *gaps between expectation and perception*).It was excluded from the top five list of the survey whereas appearing in both CS and QA interview top results. This might again relate to the participation of general users in the survey who care less about the gaps.

Consensus/Divergence Among the Stakeholder Groups

Group consensus/divergence among the interview participants. Consensus and divergence in perception of criteria

DOI: 10.1002/asi

TABLE 5. Interview participants' inter-group consensus/divergence on criteria importance perceptions.*

DL levels	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;** ; efficiency; security; cost **interoperability/compatibility**	**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; personalization	**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 complianc *e*;** organizational accessibility; compatibility; network effect	**Productivity of community members; outcome against predetermined institutional goals**

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

importance have been identified among the three stakeholder groups from the interview. Table 5 lists the consensus and divergence criteria from the CS results. The reason for using CS instead of QA results was that the criteria were identical among interviewees in CS, and thus, ready for comparison. In contrast, there was too much variance in QA on the criteria being mentioned.

The determination 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 criterion with the sum value larger or smaller by a factor of 2 over any of the other two groups, then this criterion was considered as having a divergent inter-group ranking. Otherwise, it was considered consensus. For example, the sum of the ranking value for content *usefulness to target users* is 15, 5, and 13, respectively, for the administrator, developer and researcher 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.

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

Group consensus/divergence among the survey participants. Not all DL evaluation criteria included in the survey have statistically significant differences among the five DL stakeholder groups. ANOVA results show that only 11 out of the 51 criteria (22%) have statistically significant inter-group differences on the criteria importance ratings. Table 6, a summary of the ANOVA results, demonstrates that service, interface, and user evaluation criteria received more consensus among the groups on the importance ratings, which is in line with the interview results. In contrast, the context evaluation criteria had the most group divergence.

Scheffe's post-hoc test results showed that the differences existed only among some of the five stakeholder groups. Furthermore, the differences existed primarily between the 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, or researchers highly perceived was sometimes the ones 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 held the opposite view. C*omprehensiveness* of collection was the only criterion that had higher rankings from the users. Interestingly, no significant effect was found between the developers and any of the other four groups.

In addition to the statistically significant effects, group differences can also be found through comparing the top-ranking criteria among the stakeholder groups. Some criteria are on the top five lists from all stakeholder groups (see text in bold in Table 7), while the others are perceived as being important by some of the groups. For instance, content evaluation had three criteria (i.e., *accessibility, accuracy,* and *usefulness*) that received all five groups' importance perceptions. 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. Additionally, *comprehensiveness* and *fidelity of information* only showed up in the users' and developers' top five lists, respectively. The succeeding holistic DL evaluation model section has more elaboration on inter-group consensus and divergence.

Clearly, the service evaluation had the largest inter-group consensus (100%), and the technology evaluation received the least agreement (29%) with respect to the five top 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. Lower agreement for the technology evaluation was also found in the interviews. The underlying reason might be associated with the unfamiliarity

DOI: 10.1002/asi

TABLE 6. Statistically significant inter-group divergence among survey participants ($n = 431$).

DL levels	Criteria	ANOVA results	Groups with sig. difference (mean difference[a], α)
Content	Appropriateness to target users	$F(4,423) = 3.889,\ p < 0.005$	Administrator -user (.78, .05)
	Comprehensiveness	$F(4, 425) = 5.048,\ p < 0.001$	Librarian – user (−.53, .005)
Technology	Appropriateness to digital information	$F(4,410) = 4.136,\ p < 0.005$	Administrator -user (.80, .05); librarian – user (.46, .05)
	Interoperability	$F(4,415) = 4.042,\ p < 0.005$	Librarian – user (.47, .05)
	Security	$F(4,423) = 3.618,\ p < 0.01$	Administrator-user (.84, .05)
Interface	Appropriateness to target users	$F(4,424) = 8.116,\ p < 0.001$	Administrator-user (.95, .005); librarian–user (.54, .001); researcher-user (.72, .005)
User	Acceptance	$F(4,421) = 3.991,\ p < 0.005$	Librarian–user (.42, 05)
Context	Copyright compliance	$F(4,416) = 6.753,\ p < 0.001$	Administrator-user (1.09, .05); librarian–user (.82, .001)
	Extended social impact	$F(4,410) = 3.646,\ p < 0.005$	Researcher-user (.71, .05)
	Integrity to org. practice	$F(4,414) = 4.057,\ p < 0.005$	Librarian–user (.51, .05)
	Managerial support	$F(4,416) = 5.152,\ p < 0.001$	Administrator-user (1.00, .05); librarian–user (.45, .05)

[a]Given that the first group rated higher than the second, the mean difference is positive. Otherwise, it is negative.

with DL technology by the majority of the stakeholders except the developers.

The Proposed Holistic DL Evaluation Model

The holistic DL evaluation model was constructed by analyzing the 431 cases of the online survey data. The model contains 19 core and 18 group-based criteria. The full definitions of these criteria can be found at http://hdl.rutgers.edu/1782.2/rucore10001600001.ETD.17104. The core criteria are those with higher importance rankings and perfect consensus among the five stakeholder groups, whereas the group-based criteria are selectively extracted from a pool of important criteria with lower agreement rates. First, the group-based criteria should be those perceived important criteria that have statistically significant inter-group differences (see Table 6). For those with no significant effects according to the post-hoc results, they should meet this condition before being included in the model: They must be within the top five of a given stakeholder group (see Table 7) and on the top five list of a given DL level (see Table 4).

The holistic DL evaluation model. Figure 3 is the proposed holistic model for DL evaluation, and it comprises six sets of concentric circles. Each set contains important criteria at a given DL level: the context at the top reflects the highest DL level, the content and technology at the bottom represent the two fundamental DL components, and the interface in the middle demonstrates 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.

Within a concentric circle, whereas the criteria in the center are core criteria with consensus from all the five stakeholder groups, the ones in the radiated outer rings are group-based criteria mapping the five various groups' interests. The key at the right bottom denotes the stakeholder group representations, including (USR) for general user, (RES) for researcher, (LIB) for librarian, (DEV) for developer, and (ADM) for administrator. Each outer ring contains a 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 outer rings, the more inter-group divergence a given DL level has regarding what should be evaluated at the level. For instance, the content circle has five outer rings with five different criteria, and the service circle has no outer rings. This is associated with the fact that important service level evaluation criteria reached 100% inter-group consensus, whereas the most divergence was found for important content evaluation criteria.

The distance of the outer rings from the centers represents the degree of inter-group consensus. The closer to the centers, the more agreements were reached among the stakeholder groups. Taking the Content concentric circle for instance, *comprehensive*, *integrity*, and *fidelity* were important to only one stakeholder group for each and, therefore, stay in the farer outer rings. In contract, *ease of understanding* was significant to four out of the five stakeholder groups except the administrators and, thus, is in the closest outer ring to the center.

Further elaboration on the model. Below are further elaborations on the model, starting from the fundamental DL levels (i.e., content, technology and interface) to the higher levels (i.e., service, user, and context). The elaborations focus on (a) what criteria are included as core as opposed to

DOI: 10.1002/asi

TABLE 7. Comparison of the top five criteria among the five groups of survey participants.

DL levels	Criteria	Administrator ($n = 25$)	Developer ($n = 36$)	Librarian ($n = 160$)	Researcher ($n = 53$)	User ($n = 157$)
Content	**Accessibility**	**X**$^{+b}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Accuracy	***X***$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Usefulness	**X**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Ease of understanding		X	X	X	X
	[a]Appropriateness	X^{+}		X	X	
	[a]Comprehensiveness					X
	Fidelity		X			
	Integrity of information	X				
Technology	**Ease of use**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Reliability	X^{+}	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	[a]Interoperability	X	X	X	***X***$^{+}$	
	Effectiveness	X		X	X	X
	[a]Security	X^{+}	X^{+}	X^{+}		
	Efficiency		X		X	X^{+}
	Display quality					X
Interface	**Ease of use**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Effectiveness	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Consistency	**X**	**X**$^{+}$	**X**	**X**	**X**
	[a]Appropriateness	X^{+}	X	X^{+}	X^{+}	
	Interaction support	X	X	X		X
	Effort needed				X	X^{+}
Service	**Accessibility**	**X**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Integrity	**X**	**X**	**X**	**X**	**X**
	Reliability	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Responsiveness	**X**$^{+}$	**X**	**X**	**X**	**X**
	Usefulness	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
User	**Successfulness**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Satisfaction	**X**$^{+}$	**X**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Efficiency of task complete	**X**	**X**$^{+}$	**X**	**X**	**X**$^{+}$
	Use/reuse	X	X	X		X
	[a]Acceptance	X^{+}	X^{+}	X^{+}	X	
	Productivity				X^{+}	X
Context	**Sustainability**	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	Collaboration/sharing	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$	**X**$^{+}$
	[a]**Managerial support**	**X**$^{+}$	**X**	**X**	**X**$^{+}$	**X**
	[a]Copyright compliance	X	X^{+}	X^{+}	X	
	Network effect	X				X^{+}
	Outcome		**X**	X		
	[a]Extended social impact				X	
	Productivity					X

[a]Criteria statistically proven to have an inter-group difference.
[b]X^{+} denotes that the criterion is within the top three of a given stakeholder group.

group-based and (b) what implications these criteria hold for DL evaluation.

- 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. The model suggests that all digital content should be evaluated in terms of the extent to which they are readily *accessible*, *accurate* without noticeable errors, and *useful to target users* in achieving certain goals. It also implies that digital content evaluation could be tailored by adopting the group-based criteria in the outer rings if knowing who would 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* and *appropriateness* should be highlighted.

An ideal evaluation should include both core and group-based criteria in the model. However, there is frequently a restriction on the number of criteria included. If this is the case, the group-based criteria could serve as a basis for selection.

Compared with the criteria for the other levels of evaluation, the criteria at the content level have larger inter-group variance. Except for the researcher and librarian groups, whose criteria (i.e., *ease of understanding* and *appropriateness*) are shared with some other groups, the remaining three groups have their own unique criteria, including

DOI: 10.1002/asi

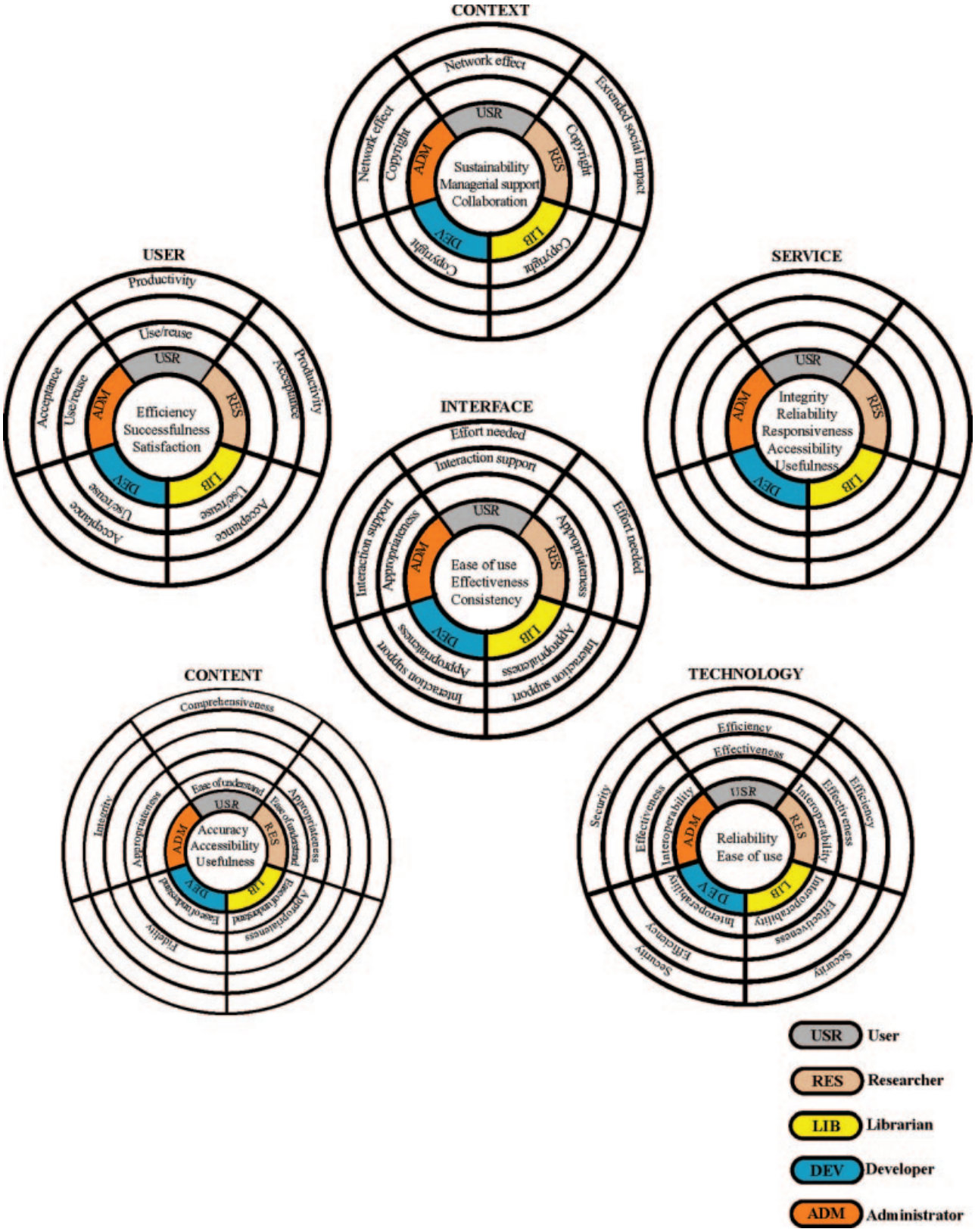

FIG. 3. The proposed holistic DL evaluation model.

DOI: 10.1002/asi

comprehensiveness of collection from the general users, *integrity* from the administrators, and *information fidelity* from the developers.

- Technology Level Evaluation Criteria

The Technology concentric circle (the bottom right) summarizes the important criteria for evaluations of digital technology, including hardware and software. In total, there are four group-based criteria and two core criteria. The model suggests that *reliability* and *ease of use* are the two important criteria to all five stakeholder groups and, thus, should be addressed in every digital technology evaluation. Meanwhile, although in the outer rings, *interoperability* and *effectiveness* might also deserve serious attention, because they both have received consensus from a large percentage of the groups (four out of five). Unlike the ones for content evaluation, all group-based technology evaluation criteria share agreement among more than three stakeholder groups.

Following the group-based criteria selection rules (see the beginning of the The Proposed Holistic DL Evaluation Model section), only two users' group-based 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', developers', and users' outer rings. Likely, the user and researcher groups have more agreement than among the remaining groups. The only exception is *interoperability*, which is on the researchers' top list but not the users'.

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. While the developers tend to opt for *efficiency*, the administrators and librarians are more concerned with *effectiveness.*

- 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, there are three group-based criteria and three core criteria at this level. The three core criteria are *ease of use*, *consistency*, and *effectiveness*. In addition to the core criteria, two group-based criteria, s*upportiveness to HCI* and *appropriateness to target users*, also have high inter-groups consensus (four out of five). Furthermore, general users tend not to think much about the *appropriateness* to them, whereas the other four groups are more sensitive to this issue. This finding is similar to the content *appropriateness*. In contrast, researchers tend to give a higher degree of importance to the criterion 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 *effort needed* as an important criterion for digital interface evaluation. Actually, they are the only two groups that include the criterion in their top lists. Again, the three groups from the librarianship professional domain tend to agree the most in their perspectives. Unanimously, the three regard *interaction support, appropriateness, ease of use, effectiveness,* and *consistency* as the top DL interface evaluation criteria.

- 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, tutorial, and term suggestion. The most noticeable feature in the circle is the blank outer rings, which indicates no group-based criteria. Throughout the entire study, all stakeholder groups agreed on the five top evaluation criteria: *accessibility, reliability, responsiveness, usefulness to target users,* and *integrity to information seeking path*. Therefore, digital service evaluations should address all the five, and the outcomes of the evaluation adopting the five criteria should be able to reflect the needs of the heterogeneous stakeholder groups.

- User Level Evaluation Criteria

The User concentric circle (the middle left) indicates the important criteria for assessing DL indirectly from users' attributes. All five stakeholder groups regard *efficiency*, *successfulness,* and *satisfaction* as important criteria. Meanwhile, *acceptance* and *use/reuse* are another two criteria that are widely perceived to be important among the stakeholder groups (four out of five). Only the researchers do not include *use/reuse* in the top list, and the general users do not perceive *acceptance* as important criteria for DL evaluation at the user level.

The two non-professional librarianship groups (i.e., general users and researchers) concur on user *productivity* as being important. None of the three professional librarianship groups includes this criterion in their top list. Instead, the three share group-based criteria, *acceptance* and *use/reuse.*

- Context Level Evaluation Criteria

The Context concentric circle (the top) suggests the important criteria for assessing DL from the following two dimensions: (a) how well a given DL fits into a larger contextual (e.g., institutional, social, cultural, economic, and legal) practices; and (b) what impacts the DL has on these contextual practices. *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 an important criterion. Besides, two other group-based criteria are *network effect* with other resources and *extended social impact*.

Similar to the content evaluation, the Context level has more scattered group-based evaluation criteria. For instance, only researchers perceive *extended social impact* as being important, and only administrators and users highlight

DOI: 10.1002/asi

network effect. However, the difference between the two DL levels is that general users, librarians, and developers have only one group-based criterion for each at the context level, whereas there are at least two group-based criteria for each group at the content level. This implies that the three groups tend to care less about DL evaluation at the context level, which is different from the administrator and researcher groups' perceptions.

In sum, except with the digital service evaluation whereby all important criteria are core and no group-based criteria have been identified, DL evaluations at the other five levels have two to three core evaluation criteria and three to five group-based criteria. A DL evaluation can be conducted flexibly by adopting all core and selective group-based criteria in any of the six concentric circles based on evaluation objectives and target stakeholders' interests. However, to get a holistic picture of a DL, it is essential to assess the DL at various levels and from different standpoints via examining all core and group-based criteria in the model. Of course, the evaluation should not necessarily fit into a single study. Instead, a 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 it also contains the group-based criteria on the top lists of some stakeholder groups. Such an evaluation model is capable of reflecting different stakeholders' needs of all kinds. Second, the model comprises important evaluation criteria for DL evaluations at all six levels, 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 should be useful for examining how well a given DL is developed as a whole whereas still being flexible in conducting DL evaluations at individual levels and for specific stakeholder groups.

Verification of the Holistic DL Evaluation Model

The most and least significant criteria. In general, there was a high replicability between the *confirmation* and *verification* stages. Among the 27 top ranked criteria from the experiment, more than three-fourths (21, 78%) were also on the top lists in the survey: *usefulness, accessibility,* and *accuracy* for content; *efficiency, reliability, interoperability* and *ease of use* for technology; *effectiveness, ease of use, effort needed* and *consistency* for interface; *reliability, usefulness, responsiveness,* and *integrity* for service; *successfulness, efficiency of task completion, use/reuse* and *satisfaction* for user; and *sustainability* and *copyright compliance* for context. Four of the remaining six criteria (i.e., content *ease of understanding*, technology *security*, interface *supportiveness of human-computer interaction*, and user *acceptance*), although not on the top five from the survey, were all ranked as sixth out of eight or nine criteria at the corresponding DL levels. The only two left behind were content *comprehensiveness* and *extended social impact* for context, which were perceived to be important by only the user group for the former and the researcher group for the latter in the survey.

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. Again, digital service criteria received the highest consistency between the two stages. All top five criteria and the least criterion remained the same.

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, one-way ANOVA test showed that the statistically significant inter-group divergence did not hold in the verification stage for all the criteria from the survey results (Table 6), except *comprehensiveness* ($F(4, 25) = 6.174,\ p < 0.001$). Presumably, this is because in the experiment setting, in spite of the difference of stakeholder group affiliation, the participants all searched the university Web for similar tasks. The similar task patterns might have an impact on identifying true group differences.

Although only one criterion had a statistically significant inter-group difference, some potential inter-group divergences in the top perceived important criteria were detected at each of the six DL levels. For a given DL level, some criteria were on the top five lists of all stakeholder groups, while the others were perceived as being important only by some of the groups.

In general, a high percentage (81%) of the criteria in the proposed model has been placed correctly as either core-based or group-based 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 associated with group affiliation and five group-based criteria (technological *security and efficiency; supportiveness of HCI, use/reuse,* and *copyright compliance*) might need to be changed to core. Therefore, the consistency is more valid in the core criteria part of the proposed DL evaluation model. Eighteen out of the 19 (95%) have received consistent results from both stages.

As for the group-based criteria, some changes were observed regarding which stakeholder groups perceived which criteria to be important. The original divergence between the general user group and the other four groups did not hold in the experiment settings. Similarly, no observations were received to support earlier findings regarding the frequently shared perspectives among the three groups from the librarianship domain. This again might be associated with the similar search tasks among the groups.

Despite the inconsistent results for some group-based criteria, because no other criteria except content *comprehensiveness* was proven to have a statistically significant inter-group divergence, the verification of the group-based criteria part of the model cannot be made via this experiment.

DOI: 10.1002/asi

Moreover, in spite of the plausible change implication from the experiment, no modifications should be made in the model for the time being, considering the following flaws in the experiment that might affect the validity of the experiment findings:

- The similar searching task patterns of the heterogeneous stakeholder groups.
- The inadequate sample size, especially of administrators ($n = 4$).
- The incomplete inclusion of all important context evaluation criteria in the post-search questionnaire.

Apparently, the model needs to be further tested in more diverse DL-use settings with more stakeholders' involvements, whereas the design drawbacks need to be eliminated. Nevertheless, considering the consistent results across the stages regarding the top/least perceived criteria at the six DL levels and among the DL groups, it can be claimed that the proposed holistic DL evaluation model has been essentially verified by the experiment, especially at the content, interface, and service levels.

Further Discussions on the Research

The research is proved effective and significant from the following aspects: (a) consistently perceived important criteria across the three research stages, (b) reliable inter-group divergence on criteria importance perceptions, especially the divergence between the user and the other stakeholder groups, and finally (c) 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 heterogeneous DL stakeholder groups, and adopting the promising framework proposed by DL experts. Despite a few weaknesses in the *verification* stage research design, the research findings are valuable to ongoing DL innovations. The section will further discuss integrated findings across the three research stages with an emphasis on 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.

The Needs of Prioritizing DL Research and Developments

Throughout the three research stages, the majority of most/least perceived important criteria have been consistently identified for DL evaluations at the six levels. In general, what these DL stakeholders are concerned about is being able to access high-quality content and service (the Premise; e.g., content *accessibility* and service *sustainability*), Their second concern is the 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 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 (e.g., *behavior change, extended social impact*) or non-core processes and premises (e.g. *personalization, courtesy of a service, conciseness of information*). These findings also suggest a caution in the current trend in personalization and customization of information retrieval systems, including DLs. When designing a personalizable DL, one has to bear those criteria with higher priorities (e.g., ease of use and effort-needed) in mind.

Similarly, when searching a DL for their tasks, the stakeholders were consistently concerned about the DL levels with which they were directly interacting: content, technology, interface, and service. In general, contextual factors received lower priorities, except for *sustainability* of a DL, a factor that has a large impact on content and service access and, hopefully, could be partially addressed through preservation—currently one of the hot topics. The preference difference also impacts importance ratings on a few criteria across the research stages. For instance, *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. This may be because that the interview did not include general users and librarians as informants who tended to place their emphasis more on the direct effects of DL use.

In general, consistently perceived important and unimportant criteria across the research stages suggest a need and feasibility for prioritization in DL research and development. DL researchers and developers should first make DL contents, technologies, and services readily 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.

Divergence among DL Stakeholders

The research consistently identifies a divergence among the stakeholder groups regarding what criteria should be used for DL evaluation, which is in accordance with Marchionini's multifaceted framework (2000). In exploration and confirmation research stages, the 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 on the technology evaluation is presumably associated with the unfamiliarity with the level to the majority of the stakeholders except the developers. Meanwhile, complexity of content (i.e., the mixture of evaluation objects in terms of meta-information, information, and collection) and indirect relationship between DL use and context might be the two factors causing the larger divergence for content and context evaluation criteria.

Additionally, not all but a small portion of the criteria has a statistically significant inter-group difference. This suggests the feasibility of conducting a generic DL evaluation embracing multiple viewpoints from various stakeholders, whereas it is necessary to tailor DL evaluations to meet

DOI: 10.1002/asi

diverse preferences from heterogeneous stakeholders. Further, the survey finding regarding the fact that the three professional librarianship groups (i.e., administrators, developers and librarians) tend to have more agreements on criteria importance than the other non-professional librarianship groups verifies the divergence proposition pinpointed by Borgman (1999) and Saracevic (2000).

Different Voices From DL Users

The research also 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 digital content, technology, and interface evaluations are not favored by the general users. Instead, DL 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 are in line with Xie's arguments (2006, 2008). Considering that in reality the other than user groups comprise the key players in DL innovation and there is very little users' involvement in DL development according to the interview findings, the research outcome provides an alert to DL researchers and professionals about the different voices from DL end users.

New Evaluation Criteria Augmenting to the Existing Research Body

By comparing the proposed holistic DL model with the literature review findings in the exploration stage, one might see that the existing DL evaluations essentially embrace the important criteria highly perceived by the various stakeholder groups. This is especially true for the content, interface, service, and user level evaluations. Table 8 shows the criteria, suggested by the model, that have been or have not been (with the blank right columns) adopted in the previous studies.

Among the 37 important criteria from the proposed model, including the core and the group-based criteria, 13 (35%) criteria have not yet been examined in 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) via their evaluation on 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 evaluation studies have been found to address the issue. This again supports Saracevic's (2000) and Bearman's (2007) assertion that contextual effects of DL have not been adequately investigated. Technology level evaluation is another weak area. One core (i.e., *ease of use*) and two group-based (i.e., *interoperability* and *security*) criteria have not yet been used in any DL evaluations, in spite of the suggestion in Kwak et al.'s (2002) framework for examining the *security* issue.

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

Two factors are associated with the gaps. Firstly, there might be a lack of awareness of the importance of those criteria, which are not directly associated with DL use, such as *collaboration/sharing* in DL development and application, and *extended social effect* in terms of how DLs change our daily lives, norms, cultural exchanges, etc. Secondly, it might be difficult to develop a valid instrument to measure a given criterion. For instance, it might not be practical to evaluate content *comprehensiveness* and *integrity* to other resources, because there is hardly a way of examining 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 research is needed to study these overlooked important criteria and to develop valid assessment instruments to measure them.

Among the 12 criteria with no previous adoption, several tend to be more DL specific. For instance, *sustainability* tends to be more crucial in DL settings, considering enormous human, financial, and technological resource investments. Similarly, *collaboration/sharing* deserves more attention in DL environment. Additionally, *extended social effects* should be highlighted because of DLs' broader applicable and influenced areas. Therefore, this study is able to fill a gap in DL innovation where there are lacks of DL-specific criteria for evaluations.

The Validity and Value of the Proposed Holistic DL Evaluation Model

The size of the DL criteria pool was reduced from original 90 from the literature review to 77 from the interviews, and eventually 37 important criteria in the proposed holistic DL evaluation model after the large-scale survey. The model should be able to serve as one of the most comprehensive models for DL evaluation for the following reasons: (a) being in light of two promising conceptual DL evaluation frameworks: Marchionini's (2000, 2003) multifaceted approach and Saracevic's (2000) stratified model; (b) being grounded on the perspectives of heterogeneous stakeholder groups, including administrator, researcher, developer, librarian, and general user; (c) covering all DL levels from content, technology, interface, service, and user to context; (d) relying on triangulation methods (i.e., interview, online survey, and experiment) from the three stages of which the research purposes and instruments are interrelated and interdependent and the results are complementary; and (e) being derived from the consistent results across the research stages.

Theoretically, this study is able to contribute to the DL research body in the following two ways: (a) This study further examines the divergence among various DL stakeholders

DOI: 10.1002/asi

TABLE 8. The adoption status of the important criteria in the existing studies.

DL level	Criteria in the model	Existing evaluation studies with the criteria adopted
Content	Accessibility	Adams & Blandford, 2001; Bishop, 1998; Jones et al., 1999; Wilson & Landoni, 2001
	Accuracy	Bergmark et al., 2002; Jones et al., 1999; Marchionini, Plaisant, & Komludi, 2003; Zhang, Low, Smoliar, & Wu, 1995
	Usefulness	Zhang & Li, 2008; Tsakonas & Papatheodorou, 2008
	Appropriateness	Borgman, Leazer, Gilliland-Swetland, & Gazan, 2001; Ding, Marchionini, & Soergel, 1999
	Ease of understanding	Khoo, Devaul, & Sumner, 2002, Zhang, 2004
	Fidelity	Jones, Gay, & Rieger, 1999; Kenney et al., 1998
	Comprehensiveness	
	Integrity	
Technology	Reliability	Champeny et al., 2004; Papadakis, Andreou, & Chrissikopoulos, 2002
	Effectiveness	Bosman, Bruza, Van de Weide, & Weusten, 1998; Hee et al., 1999; Jones & Lam-Adesina, 2002; Khoo 2001; Larsen, 2000; Rui, Gupta, & Acero, 2000; Salampasis & Diamantaras, 2002; Sanderson & Crestani, 1998
	Efficiency	Fuhr, Klas, Schaefer, & Mutschke, 2002; Kengeri et al., 1999; Larsen, 2000; Xie, & Wolfram, 2002
	Ease of use	
	Interoperability	
	Security	
Interface	Ease of use	Champeny et al., 2004; Hill et al., 2000; Huxley, 2002; Khoo et al., 2002; Papadakis et al., 2002; Zhang et al., 2008
	Effectiveness	Browne & Gurrin 2001; Park, 2000; Zhang, Li, Liu, & Zhang, 2008
	Consistency	Salampasis & Diamantaras, 2002; Wesson & Greunen, 2002; Zhang, 2004
	Appropriateness	Zhang, 2004
	Effort needed	Larsen, 2000; Zhang, 2004; Zhang et al., 2008
	Interaction support	Peng et al., 2004
User	Successfulness	Wildemuth et al., 2003; Zhang, 2004
	Satisfaction	Bishop et al., 2000; Bollen & Luce, 2002; Cullen, 2001; Wilson & Landoni, 2001
	Efficiency	Jones & Lam-Adesina, 2002; Larsen, 2000; Meyyappan, Foo, & Chowdhury, 2004; Shim, 2000
	Acceptance	Bollen & Luce, 2002; Mead & Gay, 1995; Thong, Hong, & Tam, 2002
	Use/reuse	Abbas, Norris, & Soloway, 2002; Bishop, 1998; Bollen & Luce, 2002; Borghuis et al., 1996; Brophy, Clarke, Brinkley, Mundt, & Poll, 2000; Carter & Jones, 2000; Cullen, 2001; Entlich et al., 1996; Hauptmann & Jin, 2001; Jones, Cunningham, Mcnab, & Boddie, 2000; Lankes et al., 2003; Larsen, 2000; Marchionini, 2000; Shim, 2000; Sumner & Dawe, 2001
	Productivity	
Service	Accessibility	Lankes et al., 2003; Cullen, 2001
	Reliability	Cullen, 2001; Shachaf, Oltman, & Horowitz, 2008;
	Responsiveness	Cullen, 2001; Lankes et al., 2003; Shachaf et al., 2008; White, 2001
	Integrity	
	Usefulness	
Context	Copyright compliance	Jones et al., 1999
	Social impact	Places et al., 2007
	Sustainability	
	Managerial support	
	Collaboration	
	Network effect	

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 tend to have more in common in their perceptions; and (b) this study generates a comprehensive framework for benchmarking DL evaluations towards various directions and for different purposes. Therefore, it can likely fill gaps in current DL research area, especially where little is known about what kinds of differences exist among various DL stakeholders in perceiving DLs and how evaluation should be effectively conducted by soliciting diverse stakeholders' input and reflecting more DL specific characteristics. In general, the two aspects of the contribution support the earlier proposition of multifaceted and multilevel evaluations proposed by many researchers (Harter & Hert, 1997; Nicholson, 2004; Marchionini, 2000; O'Day & Nardi, 2003; Saracevic, 2000).

Pragmatically, the proposed holistic evaluation model can provide DL developers and assessors with a comprehensive and flexible toolkit for conducting systematic designs and evaluations. Using the toolkit, the developers or assessors can readily conduct tailored DL evaluations for various purposes and with multiple perspectives. Specifically, as suggested by the model, DL evaluation at a given level could be conducted by adopting all its core criteria and selecting some group-based criteria based upon evaluation objectives and target stakeholders' interests. The Further Elaboration on the Model section has detailed "how to" implications.

DOI: 10.1002/asi

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

Limitation and Future Research

The proposed holistic DL evaluation model has a few limitations. First, the general users are all from university settings. The limit might affect the representativeness of the model. For example, the divergence between general users and the other four stakeholder groups as suggested by the model might change in other than academic settings because non-academic users (e.g., children, elderly) might have different perceptions on important evaluation criteria. With more and more DLs being developed for less sophisticated users, it becomes extremely crucial to listen to these users' voices. Additionally, the model excludes DL funders' perception, and accordingly its comprehensiveness might be affected. Second, the model construction heavily relies on stakeholders' subjective thoughts and might have negative impacts on the validity of the framework. Third, the model has been tested with a single type of DL: an academic library Web site. Tests with other types of DLs might have different results. Last, the model describes what criteria should be used for various levels of DL evaluation. But it does not specify how to apply appropriate criteria into actual evaluation studies.

Although DLs are growing in numbers in the past couple of years and have advanced in technologies, the general components and boundaries of them remain almost unchanged. Bearman (2007) pointed that digital library was a "mature information service application." Additionally, according to ISI Web of Knowledge, the average Cited-Half-Life for library and information science journal articles is 6.9, which means that 50% of the total citations from the current year are dated back to the past 7 years. Therefore, although the studies were conducted in 2005 and 2006, the major research findings should still be applicable to current DLs, useful to future DL research, and worthy to be tested in current DLs.

Accordingly, further studies are needed to overcome these weaknesses, especially in the following areas: (a) enriching the model by the inclusion of more heterogeneous stakeholders' opinions; (b) testing the model in various and, presumably, even beyond academic DL settings; and (c) developing a methodological framework for supporting the operationalization of these criteria and empowering the flexibility of conducting various tailored evaluations in light of the holistic model. Additionally, it would be helpful to develop exemplar evaluation instances with demonstrations on how to select appropriate group-based criteria to achieve specific goals.

Conclusion

The article presents a holistic model for DL evaluation, a model that was constructed through three incremental phases of work: *exploration* (a literature review and interviews with card sorting), *confirmation* (a large-scale survey), and *verification* (an evaluation of an extant DL). Based upon a series of examination of heterogeneous stakeholder groups' viewpoints about DL evaluations at various levels, the holistic model outlines specific criteria that should be used and could be tailored for multifaceted and multilevel DL evaluations.

In general, DL stakeholders are most concerned about content *accessibility* and service *sustainability*. Their second concern is factors with direct impact on their interaction with the content and service (e.g., *ease of use, effort-needed, interoperability,* service *responsiveness*), and, third, they 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 (e.g., *behavior change, extended social impact*) or non-core processes and premises (e.g. *personalization, courtesy of a service, conciseness of information*).

Additionally, as the model suggests, DL stakeholders share a large portion of important criteria for evaluations but have a small number of criteria with inter-group divergence. Specifically, three groups of stakeholders from the professional domain (i.e., administrators, developers, and librarians) have similar perspectives, whereas general users have larger divergence from the other groups. These research findings are in accordance with some earlier studies (Borgman, 1999; Saracevic, 2000; Van House, Butler, & Schiff, 1995; Xie, 2006, 2008). Furthermore, the research reveals that despite the majority of the important criteria in the proposed model have been adopted in previous studies, a few (e.g., content *comprehensiveness* and *integrity*, technological *interoperability* and *security*, service *integrity* and *usefulness*, user *productivity*, and *sustainability*) so far have received little attention.

Overall, the research makes unique contributions to DL innovations. The proposed model fills two lacunae in the DL domain: (a) the lack of a comprehensive and flexible framework to guide and benchmark evaluations and (b) the uncertainty about what divergence exists among heterogeneous DL stakeholder groups.

Acknowledgment

This article derives from dissertation research. I want to give special thanks to my adviser, Dr.Tefko Saracevic, whose outstanding advisorship and mentorship inspired me throughout the dissertation research. I also want to thank my committee members, Drs. Heting Chu, Michael Lesk,

DOI: 10.1002/asi

Claire McInerney, and Daniel O'Connor, for their insightful comments and constructive criticism on the research.

References

Abbas, J. (2002). Middle school children's use of the ARTEMIS Digital Library. In G. Marchionini & W.R. Hersh (Eds.), Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 98–105). New York: ACM Press.

Adams A., & Blandford, A. (2001). Digital libraries in a clinical setting: Friend or foe? In P. S. Constantopoulos & I. Sølvberg (Eds.), Research and advanced technology for digital libraries. Proceedings of Fifth European Conference (ECDL 2001) (pp. 214–224). Berlin, Germany: Springer.

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

Basefky, S. (2009). The end of institutional repositories & the beginning of social academic research service: An enhanced role for libraries. Retrieved on June 17, 2009, from http://www.llrx.com/node/2177/print

Bearman, D. (2007). Digital libraries. Annual Review of Information Science and Technology, 41, 223–272.

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). Focused crawls, tunneling, and digital libraries. In M. Agosti & C. Thanos (Eds.), Research and Advanced Technology for Digital Libraries: Proceedings of the Sixth European Conference (ECDL '02) (pp. 91–106). Berlin, Germany: Springer.

Bishop, A.P. (1998). Logins and bailouts: Measuring access, use, and success in digital libraries. The Journal of Electronic Publishing, 4(2). Retrieved September 29, 2009, from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0004.207

Bishop, A.P. (1999). Making digital libraries go: comparing use across genres. In N. Rowe & E.A. Fox (Eds.), Proceedings of the Fourth ACM Conference on Digital Libraries (pp. 94–103). New York: ACM Press.

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.

Blixrud, J.C. (2002). Measures for electronic use: The ARL E-Metrics project. Retrieved September 29, 2009, from http://www.lboro.co.uk/departments/dis/lisu/downloads/statsimpracticepdfs/blixrud.pdf

Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. D-Lib Magazine, 8(6). Retrieved September 29, 2009, from http://www.dlib.org/dlib/june02/bollen/06bollen.html

Borghuis, M., Elsevier Science, Carnegie-Mellon University, Cornell University, Georgia Institute of Technology, Massachusetts Institute of Technology, et al. (1996). Tulip: Final report. New York: Elsevier Science.

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

Borgman, C.L. (2002). Final report to the National Science Foundation. In Proceedings of the Fourth DELOS Workshop, evaluation of digital libraries: Testbeds, measurements, and metrics. Retrieved September 29, 2009, from http://www.sztaki.hu/conferences/deval/presentations/final_report.html

Borgman, C.L., Leazer, G.H., Gilliland-Swetland, A.J., & Gazan, R. (2001). Iterative design and evaluation of a geographic digital library for university students: A case study of the Alexander Digital Earth Prototype (ADEPT). In P.S. Constantopoulos & I. Sølvberg (Eds.), Research and advanced technology for digital libraries: Proceedings of the Fifth European Conference (pp. 390–401). Berlin, Germany: Springer.

Bosman, F.J.M., Bruza, P.D., Van de Weide, T.P., & Weusten, L.V.M. (1998). Documentation, cataloging, and query by navigation: A practical and sound approach. In C. Nikolaou & C. Stephanidis (Eds.), Research and advanced technology for digital libraries: Proceedings of the Second European Conference (pp. 459–478). Berlin, Germany: Springer.

Brophy, P., Clarke, Z., Brinkley, M., Mundt, S., & Poll, R. (2000). EQUINOX: Library performance measurement and equality management system-Performance indicators for electronic library services. Retrieved September 29, 2009, from http://equinox.dcu.ie/reports/pilist.html

Browne, P., & Gurrin, C. (2001). Dublin City University video track experiments for TREC 2001. Retrieved September 29, 2009, from http://trec.nist.gov/pubs/trec10/papers/dcu_trec01_final.pdf

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., Gilliland-Swetland, A.J., Millwood, K.A., D'Avolio, L. et al. (2004). Developing a digital learning environment: An evaluation of design and implementation processes. In H.C. Chen, H. Wactlar, C.C., Chen, E.P. Lim, & M. Christel (Eds.), Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (pp. 37–46). New York: ACM Press.

Chowdhury, G.G., Chowdhury, S. (2003). Introduction to digital libraries. London: Facet.

Computer Science & Telecommunication Board, National Research Council (1998, July/August). Design and evaluation. A review of the state-of-the-art. D-Lib Magazine, 4. Retrieved September 29, 2009, from http://www.dlib.org/dlib/july98/nrc/07nrc.html

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

Dillon, A. (1999). TIME-A multi-level framework for the design and evaluation of digital libraries. International Journal of Digital Libraries, 2 (2/3), 170–177.

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, 85–93.

Entlich, R., Garson, L., Lesk, M., Normore, L., Olsen, J., & Weibel, S. (1996). Testing a digital library: User response to the CORE project. Library Hi Tech, 14(4), 99–118.

Fuhr, N., Hansen, P., Mabe, M., & Micsik, A. (2001). Digital libraries: A generic classification and evaluation scheme. In P.S. Constantopoulos & I. Sølvberg (Eds.), Research and advanced technology for digital libraries: Proceedings of the Fifth European Conference (pp. 187–199). Berlin, Germany: Springer.

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, Paris, France, 597–612.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet the assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42(2), 237–288.

Goncalves, M.A., Moreira, B.L., Fox, E.A., & Watson, L.T. (2007). "What is a good digital library?" A quality model for digital libraries. Information Processing and Management, 43(5), 1416–1437.

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

Harter, S.P., & 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 (pp. 3–94). Medford, NJ: Information Today.

Hauptmann, A., & Jin, R. (2001). Video retrieval with the Informedia Digital Video System. Retrieved September 29, 2009, from http://trec.nist.gov/pubs/trec10/papers/CMU-VideoTrack.pdf

Hee, M., Yoon, Y.I., & Kim, K.C. (1999). Unified video retrieval systems supporting similarity retrieval. In T.J.M. Bench-Capon, G. Soda, & A.M. Tjoa (Eds.), Proceedings of the Tenth International Workshop on Database and Expert System Applications (pp. 884–888). New York: Springer.

Hill, L.L., Carver, L., & Larsgaard, M., Dolin, R., Smith, T.R., Frew, J., et al. (2000). Alexandria Digital Library: User evaluation studies and system design. Journal of the American Society for Information Science, 51(3), 246–259.

Huxley, L. (2002). Renardus: Following the Fox from project to service. In M. Agosti & C. Thanos (Eds.), Research and advanced technology for digital libraries: Proceedings of the Sixth European Conference (pp. 157–171). Berlin, Germany: Springer.

DOI: 10.1002/asi

Isfandyari-Moghaddam, A., & Bayat, B. (2008). Digital libraries in the mirror of the literature: Issues and considerations. Electronic Library, 26(6), 844–862.

Jones, G.J.F., & Lam-Adesina, A.M. (2002). An investigation of mixed-media information retrieval. Research and Advanced Technology for Digital Libraries. Proceedings of the 2nd European Conference, 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 September 29, 2009, from http://dlib.org/dlib/november99/11jones.html

Jones, S., Cunningham, S.J., Mcnab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. International Journal of Digital Library, 3(2), 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.

Kengeri, R., Seals, C.D., Reddy, H.P., Harley, H.D., & Fox, E.A. (1999). Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, and NDLTD. International 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. In C. Nikolaou & C. Stephanidis (Eds.), Research and advanced technology for digital libraries: Proceedings of the Second European Conference (pp. 279–293). Berlin, Germany: Springer.

Khalil, M.A., & Jayatilleke, R. (2000). Digital libraries: Their usage for the end user point of view. Proceedings of the National Online Meeting, New York, NY, 179–187.

Khoo, M. (2001). Ethnography, evaluation, and design as integrated strategies: A case study from WES. In P.S. Constantopoulos & I. Sølvberg (Eds.), Research and advanced technology for digital libraries: Proceedings of the Fifth European Conference (pp. 263–274). Berlin, Germany: Springer.

Khoo, M., Devaul, H., & Sumner, T. (2002). Functional requirements for online tools to support community-led collections building. In M. Agosti & C. Thanos (Eds.), Research and advanced technology for digital libraries: Proceedings of the Sixth European Conference (pp. 190–203). Berlin, Germany: Springer.

Kwak, B.H., Jun, W., & Gruenwald, L. (2002). A study on the evaluation model for university libraries in digital environments. In M. Agosti & C. Thanos (Eds.), Research and advanced technology for digital libraries: Proceedings of the Sixth European Conference (pp. 204–217). Berlin, Germany: Springer.

Kyrillidou, M., & Giersch, S. (2005). Developing the DigiQUAL protocol for digital library evaluation. In M. Marlino, T. Summer, & F. Shipman (Eds.), Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 172–173). New York: ACM Press.

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 April 13, 2004, from http://www.dlib.org/metrics/public/papers/The_Dlib_Test_Suite_and_Metrics.pdf

Lindlof, T.R. (1995). Qualitative communication research methods. Thousand Oaks, CA: Sage.

Lynch, C. (2003). Colliding with the real world: Heresies and unexplored questions about audience, economics, and control of digital libraries. In A.P. Bishop, A.A. Van House, & B.P. Buttenfield (Eds.), Digital library use: Social practice in design and evaluation (pp. 191–216). Cambridge MA: The MIT Press.

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 P. Ann Biship et al. (Eds.), Digital library use: Social practice in design and evaluation (pp.119–160). Cambridge MA: The MIT Press.

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., 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.

Nielsen, J. (1993). Usability engineering. Boston MA: Academic Press.

O'Day, V.L., & Nardi, B.A. (2003). An ecological perspective on digital libraries. In A.P. Bishop, A.A. Van House, & B.P. Buttenfield (Eds.), Digital library use: Social practice in design and evaluation (pp. 65–82). Cambridge MA: The MIT Press.

Papadakis, I., Andreou, I., & Chrissikopoulos, V. (2002). Interactive search results. In M. Agosti & C. Thanos (Eds.), Research and advanced technology for digital libraries: Proceedings of the Sixth European Conference (pp. 448–462). Berlin, Germany: Springer.

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.

Places, A.S., Brisaboa, N.R., & Farina, A., Luaces, M.R., Parama, J.R., & Penabad, M.R. (2007). The Galician virtual library. Online Information Review, 31(3), 333–352.

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

Rui, Y., Gupta, A., & Acero, A. (2000). Automatically extracting highlights for TV baseball programs. In S. Ghandeharizadeh, S.F. Chang, S. Fischer, J.A. Konstan, & N. Nahrstedt (Eds.), Proceedings of the Eight ACM International Conference on Multimedia (pp. 105–115). New York: ACM Press.

Salampasis, M., & Diamantaras, K.I. (2002). Experimental user-centered evaluation of an open hypermedia system and Web information seeking environments. Journal of Digital Information, 2(4). Retrieved September 29, 2009, from http://journals.tdl.org/jodi/article/viewArticle/57/60

Sanderson, M., & Crestani, F. (1998). Mixing and merging for spoken document retrieval. In C. Nikolaou & C. Stephanidis (Eds.), Research and advanced technology for digital libraries: Proceedings of the Second European Conference (pp. 397–407). Berlin, Germany: Springer.

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, 33, 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. (2000). Digital library evaluation: Toward an evolution of concepts. Library Trends, 49(2), 350–369.

Shachaf, P., Oltmann, S.M., & Horowitz, S.M. (2008). Service equality in virtual reference. Journal of the American Society for Information Science and Technology, 59(4), 535–550.

Shim, W. (2000). Measuring services, resources, users, and use in the networked environment. Retrieved September 29, 2009, from http://www.arl.org/bm~doc/emetrics-2/pdf

Sumner, T., & Dawe, M. (2001). Looking at digital library usability from a reuse perspective. In E.A. Fox & C.L. Borgman (Eds.), Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries (pp. 416–425). New York: ACM Press.

Thong, J.Y.L., Hong, W.Y., & Tam, K.Y. (2002). Understanding user acceptance of digital libraries: What are the roles of interface characteristics,

DOI: 10.1002/asi

organizational context, and individual differences? International Journal of Human-Computer Studies, 57(3), 215–242.

Tsakonas, G., & Papatheodorou, C. (2008). Exploring usefulness and usability in the evaluation of open access digital libraries. Information Processing and Management, 44(3), 1234–1250.

Van House, N.A., Butler, M., & Schiff, L. (1996). Needs assessment and evaluation of a digital environmental library: The Berkeley experience. Retrieved September 29, 2009, from http://people.ischool.berkeley.edu/~vanhouse/dl96.html

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. In P. Kotzé, L. Venter, & J. Barrow (Eds.), Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology (pp. 11–18). New York: ACM Press.

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., et al. (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, Los Alamitos, CA: IEEE. pp. 221–230.

Wilson, R., & Landoni, M. (2001). Evaluating electronic textbooks: A methodology. In P.S. Constantopoulos & I. Sølvberg (Eds.), Research and advanced technology for digital libraries: Proceedings of the Fifth European Conference (pp. 1–12). Berlin, Germany: Springer.

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(3), 433–452.

Xie, H.I. (2008). Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. Information Processing and Management, 44(3), 1346–1373.

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, CA: San Francisco, 15–24.

Zhang, Y. (2004). Moving image collections evaluation–Final report. Retrieved September 29, 2009, from http://comminfo.rutgers.edu/~miceval/progress/final_report/doc

Zhang, Y., & Li, Y.L. (2008). A user-centered functional metadata evaluation on moving image collections. Journal of American Society for Information Science and Technology, 59(8), 1331–1346.

Zhang, X.M., Li, Y.L., Liu, J.J., & Zhang, Y. (2008). Effects of interaction design in digital libraries on user interactions. Journal of Documentation, 64(3), 438–463.

DOI: 10.1002/asi