The Journal of Academic Librarianship 44 (2018) 854–863

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

Multifaceted Evaluation Criteria of Digital Libraries in Academic Settings: Similarities and Differences From Different Stakeholders

Iris Xie[a,*], Soohyung Joo[b], Krystyna K. Matusiak[c]

[a] School of Information Studies, University of Wisconsin–Milwaukee, Northwest Quad Building B, 2025 E. Newport, Milwaukee, WI 53211, United States of America
[b] School of Information Science, University of Kentucky, 320 Little Library Bldg., Lexington, KY 40506, United States of America
[c] Research Methods and Information Science Department, University of Denver, 1999 East Evans Avenue, Denver, CO 80208, United States of America

ARTICLE INFO

Keywords:
Evaluation criteria
Evaluation dimensions
Digital libraries
Academic libraries
Stakeholders

ABSTRACT

Digital library (DL) evaluation is essential to the success and enhancement of DLs. However, there is a lack of research on the assessment of comprehensive evaluation criteria across multiple dimensions of DLs. In particular, limited research is available on criteria prioritization to determine which criteria are perceived important by different stakeholders. This study was conducted to compare similarities and differences in perceptions of the importance of different DL evaluation criteria by heterogeneous stakeholders in academic settings. Ninety subjects were recruited with 30 from each of the group representing DL scholars, DL librarians, and DL users. Subjects were instructed to fill in an in-depth survey consisting of 10 evaluation dimensions with 94 criteria. ANOVA and *t*-test were applied to examine the similarities and differences among the three groups. This study reveals consensus and divergence in perceptions of criteria importance among the three groups, and indicates an inherent tension among the stakeholders. Moreover, the differences identify gaps not only between user expectations and the DL practice but also between what's desirable and what's possible in the academic environment. The findings provide a comprehensive list of criteria to guide practical evaluation of DLs, and contribute to the narrowing of the identified gaps.

Introduction

Evaluating digital libraries (DLs) is vital to their ultimate success. Through evaluation, we can recognize if the DL is effective and efficient for users, and if their needs are fulfilled. DL evaluation is also important to identify problems and issues in the DL, which can potentially create barriers to use, and ultimately defeat its purpose. Evaluation practice should be an essential component in the cycle of DL development, curation, management, and follow-up upgrades. In this study, evaluation refers to "the process of determining merit, worth or valuation of something, or the product of that process" (Scrivin, 1991, p. 139).

DLs are defined as the representations of emergent and complex forms of digital information organization and design, consisting of multiple layers and building blocks, in various stages of development (Borgman, 1999; Candela et al., 2007; Lesk, 2005; Xie & Matusiak, 2016). DLs consist of not only standalone DLs such as the New York Public Library Digital Collections, International Children's Digital Library, and Cornell University Library Digital Collections but also large scale DLs that aggregate content from individual libraries with portals for global searching and retrieval such as the Digital Public Library of America (DPLA), HathiTrust, and Europeana. DLs present a variety of resources created in the digital format as well as those converted from analog materials through digitization efforts, including print materials, images, audios, and videos (Lesk, 2005; Miller, 2011). The concepts of DLs are still evolving, corresponding to a complex notion, and cannot be captured by a simple definition (Bishop, Van House, & Buttenfield, 2003; Candela et al., 2007; Greenstein, 2000; Saracevic, 2004; Xie & Matusiak, 2016).

As DLs are complex and multifaceted in their nature, both researchers and practitioners need a set of guidelines of what to evaluate, how to measure results, when to undertake evaluation, and how to incorporate evaluation results into the DL development. Evaluation criteria used in assessing traditional library collections, although applicable to a certain extent, became insufficient for the emerging dimensions of DLs. Therefore, the evaluation of DLs requires the methods to assess those multiple aspects of DLs ranging from collection, system interface, to user needs (Carr, 2006; Kani-Zabihi, Ghinea, & Chen, 2006; Nicholson, 2004; Weiss, 2014).

Previous DL studies have contributed various evaluation criteria to examine different aspects of DLs, such as accessibility and usability of

* Corresponding author.
E-mail addresses: hiris@uwm.edu (I. Xie), soohyug.joo@uky.edu (S. Joo), krystyna.matusiak@du.edu (K.K. Matusiak).

https://doi.org/10.1016/j.acalib.2018.09.002
Received 23 April 2018; Received in revised form 31 August 2018; Accepted 13 September 2018
Available online 21 September 2018
0099-1333/

interface, user engagement, collection quality, and several others (Saracevic, 2004; Zhang, 2010). In addition, there has been concerted effort to identify different components of DLs to explain the complexity of DLs holistically. These research efforts provided a number of evaluation criteria applicable to performing the evaluation of multiple dimensions of DLs.

Yet, limited research is available on criteria prioritization to determine which criteria are perceived more or less important from different types of stakeholders. In an effort to fill these gaps in DL evaluation research, this study surveys the perceived importance of various criteria in DL evaluation from the perspectives of three groups of stakeholders, including DL users, DL librarians, and DL scholars. To avoid repetition, users, librarians and scholars are used to represent these three groups of stakeholders.

Literature review

DL evaluation research has been conducted to assess different aspects of DLs, such as comprehensive evaluation models, user-centered evaluation, outcome measures, and interface design. First of all, there were several works that suggested comprehensive DL models, which were used as the fundamental framework for DL evaluation. The DELOS DL reference model identifies six core components, including content, functionality, quality, policy, architecture, and user, in the three layers of DLs, and lays down the groundwork for explaining the relationships among the DL, the DL system, and the DL management system (Candela et al., 2007). Although the DELOS Reference Model was created about ten years ago, its framework is still considered comprehensive to conceptually define different components of a DL system. According to DELOS (Candela et al., 2007), we can infer that DLs are needed to be assessed in terms of functionality and content, rules/policy for the users and library, the quality of content and performance, services/functionality, communications, and information/content. Long-term preservation of digital content is recognized as a core function of DLs in the DELOS model as well (Candela et al., 2007). Fuhr et al. (2007) further developed a DL evaluation framework on the basis of DELOS model and constructed the specific survey for DL evaluation. Saracevic's (2004) research is another widely cited model in the area of DL evaluation. His main contribution is the identification of the comprehensive set of evaluation criteria in the six dimensions of DLs: content, technology, interface, process/service, user, and context. Zhang (2010) further advanced the Saracevic's evaluation framework by investigating the importance of those dimensions of criteria. DigiQUAL was also designed as a comprehensive tool for DL evaluation that can be useful for service quality assessment practice (Kyrillidou & Giersch, 2005). DigiQUAL involves twelve themes of service quality, such as accessibility, navigability, interoperability, collection building, resource use, and others. Noh (2010) developed a comprehensive set of evaluation criteria that covers multiple sectors of electronic resources in libraries. Her evaluation framework includes specific evaluation indicators to assess the development and use of digital resources in academic library environments.

The emphasis of the early DL research and practice has been primarily on expanding access to cultural heritage and scientific resources rather than long-term preservation and sustainability (Chowdhury, 2010; Ross, 2012; Xie & Matusiak, 2016). The evaluation frameworks developed by Saracevic (2004) and Zhang (2010) focused on the dimensions of DLs that supported effective and efficient user interaction, such as technology, interface, service, user, and context, but did not consider digital preservation. In recent years, however, preservation of born-digital and digitized content of DLs has been receiving more attention in research and practice (Beaudoin, 2012a, 2012b; Matusiak, Taylor, Newton, & Polepeddi, 2017; Miller, 2017; Rinehart, Prud'homme, & Huot, 2014). Digital preservation policy is an active area of development and evaluation (Dressler, 2017; Noonan, 2014). Chowdhury (2014) has widened the scope of DL model to cover the issue of sustainability. His conceptual model of DLs involves the three areas of DL sustainability such as economic, social, and environmental sustainability. He identified specific elements for each sustainability area, for example, funding models for economic sustainability, novel cloud-based design for environmental sustainability, and user- & context-specific design for social sustainability.

Kelly (2014) conducted a comprehensive review of professional literature about DL assessment as an effort to identify key themes in DL evaluation. She identified a range of areas of DL evaluation, such as usability evaluation, usage statistics for collection, altmetrics, DL resource reuse, and cost benefit analysis. Similarly, Petras and Stiller (2017) conducted a meta-analysis of forty-one prior studies concerning the evaluation of Europeana to explore multiple constructs, such as contexts, criteria, and measures, applicable to DL evaluation. Albertson (2015) designed an evaluation framework tailored to visual DLs, and discussed specific components to be assessed for visual resource DL systems, such as users, interactions, systems, topics, and others.

Recognizing the importance of meeting users' needs, many researchers made efforts to design user-centered DL evaluation models and conducted usability studies in DL environments. Xie (2006, 2008) proposed a user-centered evaluation framework, which posits five dimensions of evaluation criteria, including usability, collection quality, service quality, system performance efficiency, and user feedback solicitation. Jeng (2005) suggested a usability evaluation model that comprised with the four attributes of effectiveness, efficiency, satisfaction, and learnability. Joo and Lee (2011) developed a measurement instrument to assess DL usability and empirically tested its reliability and validity using structural equation modeling. The heuristic methods were also adopted to evaluate the usability of DLs. For example, Inal (2018) adopted Nielsen's (1995) web usability heuristics, and modified them to reflect the unique context of DL usability. The suggested heuristics included ten specific items, such as visibility of system status, user control and freedom, consistency and standards, error prevention, aesthetics and minimalist design, and others. Accessibility is another important area that researchers paid attention with regard to DL evaluation. Ferati and Beyene (2017) also utilized the heuristics approach and created sixteen heuristics items to evaluate the accessibility of DL. From the analysis of blind users' interaction with DLs, Xie et al. (2015, 2018) identified their help-seeking situations at the physical and cognitive levels. Those situations can be used as the assessment points to enhance the DL accessibility for blind users. Mune and Agee (2016) evaluated the accessibility of e-book collection platforms for users with physical or learning disabilities in regards to file format, layout, text adjustments, text-to-speech, and others.

DL evaluation research has also been conducted with regard to the system interface and specific components of DLs. For example, Lai, Chiu, Huang, Chen, and Huang (2014) developed an evaluation framework for system interface based on the fuzzy analytic hierarchical process (AHP) method. Using AHP, they determined the top five important evaluation criteria for DL system interface as ease-of-use, searching, language, presentation, and design. Gkoumas and Lazarinis (2015) evaluated the technical features of open source DL platforms in terms of system options, content types, support for cataloging and circulation, search options, and interoperability. Behnert and Lewandowski (2017) investigated the evaluation of information retrieval features in library systems including DLs, and suggested the methods to be useful for the evaluation of retrieval effectiveness.

Large-scale DLs with aggregated metadata and multi-layered navigation systems pose new challenges to evaluation. National Science Digital Library (NSDL) served as a testing ground for examining user interaction and quality of metadata. A number of studies were conducted with educators to evaluate the functionality of the NSDL and its support for designing learning activities (Recker, 2006; Recker et al., 2007; Xu & Recker, 2012). Bui and Park (2006) assessed the quality of metadata aggregated in NDSL in terms of consistency, completeness, accuracy, and local additions of data providers. Zavalina (2014)

examined metadata in three large-scale digital libraries and found that high-quality metadata is crucial for providing adequate subject access to rich aggregated content. Dobreva and Chowdhury (2010) conducted user-centered evaluation of Europeana, and found a gap between user perceptions and functionality. Matusiak (2017) evaluated user navigation options in a DL distributed environment using the Digital Public Library of America (DPLA) as a case study and found some users struggling with a multi-layered structure of the DPLA.

While previous research has provided different DL evaluation models, evaluation dimensions and evaluation criteria, very few studies have examined the importance of DL evaluation criteria from the key groups of stakeholders. Zhang's (2010) study is one of the few that explores the importance of DL evaluation criteria from the perspective of multiple stakeholders, but the selection of criteria is limited to the Saracevic's six dimensions.

Research questions and associated hypotheses

In order to fill the research gap, this study posed the following research question: 1) What are the similarities and differences in perceptions of the importance of different DL evaluation criteria by heterogeneous stakeholder groups?

Here is the associated hypothesis:

H1. There is no significant difference in rating each evaluation criterion of each DL evaluation dimension among the three groups of stakeholders.

A thorough document analysis was conducted as the first step to suggest an initial pool of dimensions and criteria with specific measures and data collection methods. In order to identify DL evaluation dimensions and associated criteria, the authors first conducted document analysis focusing on keywords “digital library,” “evaluation,” “criteria,” “assessment,” and other associated terms in different combinations using search operators. Google Scholar and DL related EBSCO periodical databases, such as Academic Search Complete and Library, Information Science & Technology Abstracts with Full Text were selected to search for relevant documents. The inclusion criteria are: 1) the paper covers any evaluation theories, frameworks, criteria, indicators, or measures; or 2) the paper consists of actual evaluation studies or pilot tests. Finally, eighty-five relevant documents and five DL evaluation project websites (Equinox, DigiQUAL, LibQUAL+, eVALUEd, DELOS) were chosen and further analyzed for DL evaluation dimensions and corresponding evaluation criteria.

Ten dimensions emerged from the analysis of the selected literature and the Web sites, including collection, information organization, interface design, system performance, effects on users, user engagement, services, preservation, sustainability/administration, and context of use. In addition, associated criteria identified in previous works were also incorporated into this stage of the in-depth survey. The detailed discussion of document analysis can be found in a previous published article (Joo & Xie, 2013). These DL criteria and their definitions were updated and used in the in-depth survey.

Methodology

Sampling

Ninety subjects were recruited with 30 subjects from each of the three groups: the scholars, the librarians, and the users. The scholar group includes researchers who have conducted DL research with high citations or professors who have taught DL courses. The former were identified based on the search results of Web of Knowledge or Google scholars, and the latter were selected from websites of library and information science schools. Digital librarians were randomly selected from the top 200 US colleges which have operating DLs, according to US News Rank (www.usnews.com/rankings), as well as librarians from the partner libraries. User group subjects, which consist of faculty members and students, were recruited from five partner libraries across the country. Five academic libraries were selected to be partners in the data collection stage. They represent academic libraries in different types of academic libraries, geographical locations, different volumes and designs of digital collections. The partner libraries include: University of Denver, University of Florida, University of Nevada Las Vegas, Drake University, and University of Wisconsin-Milwaukee. Their digital librarians directly participated the two-round surveys, and helped recruit user group subjects.

Table 1
Demographic information of scholar subjects (N = 30)

Demographic characteristics		Number	Percentage
Title	Professor	11	36.7%
	Associate Professor	9	30.0%
	Assistant Professor	8	26.7%
	Researcher	1	3.3%
	Professor Emeritus	1	3.3%
Gender	Male	14	46.7%
	Female	15	50.0%
	No response	1	3.3%
Age	31–40	6	20.0%
	41–50	12	40.0%
	51–60	7	23.3%
	61+	5	16.7%
Research areas	digital libraries, information retrieval, metadata, digital humanities, HCI, digital preservation, information visualization,		

Table 2
Demographic information of digital librarian participants (N = 30)

Demographic characteristics		Number	Percentage
Gender	Male	13	41.9%
	Female	18	58.1%
Age	31–40	3	9.7%
	41–50	10	32.3%
	51–60	12	38.7%
	61+	6	19.4%
Years in DL services	Average = 8.48 years (STD = 5.32) Median = 8 years Range = 6 months - 23 years		
Title	Digital Librarian, Digital Initiatives Librarian, Digital Projects Coordinator, Digital Collections Librarian, Digital Projects Manager, Head of Digital Collections, Head of Special Collections and Digital Initiatives, etc.		

Table 1 presents demographic information of 30 scholar subjects. It shows well balanced proportions by professor rank and gender. Their research areas basically include DLs, and additionally cover different aspects of DLs, such as information retrieval, metadata, HCI, and preservation. Table 2 presents demographic information of 30 digital librarian subjects. On average, they have about 8.48 (median 8) years of experiences in DL related services. Their official titles were diverse consisting of digital librarian, digital initiative librarian, and digital collection librarian. Finally, Table 3 presents demographic information of 30 user subjects. From each of the five partner institutions, four students and 2 faculty members participated in the survey. User subjects represent diverse majors of students and disciplines of faculty members.

Data collection

Two-round in-depth surveys were administered across three groups of stakeholders. Since the first round survey was quite comprehensive and in-depth, it took each subject about 40 min to complete it. In the first round, the importance of evaluation criteria was investigated based

Table 3
Demographic information of user participants ($N = 30$)

Demographic characteristics		Number	Percentage
Group	Undergraduate	8	26.7%
	Graduate	12	40%
	Faculty	10	33.3%
Gender	Male	12	40%
	Female	18	60%
Student major	Linguistics, English, Elementary and Special Education, History, Film, Curriculum and Instruction, Geographic Information Sciences, Finance, Geography, Environmental Science, Mental Health Counseling, Art, etc.		
Faculty discipline	Counselor Education, English, History, Geography, Computer Science, Law, Psychology, Journalism, Information Science		

on seven point Likert scale. In this way, importance served as a key variable to rank evaluation criteria. The purpose of the first round was to determine the importance of DL evaluation criteria under the 10 dimensions from the perspectives of different stakeholders, and moreover, to compare their similarities and differences. The 10 dimensions and 94 evaluation criteria and associated definitions were generated based on the document analysis conducted by the authors (Joo & Xie, 2013). To help subjects understand the meaning of evaluation criteria, they were provided definitions to each dimension and criterion generated by the document analysis. Also, subjects were instructed to offer additional criteria they perceived to be important, but were not part of the list. In addition, subjects were further asked in which stages of DL development and operation that each criterion should be applied. The authors created 7 phases covering the evolution of DLs: planning, prototyping, building, testing, launching, operating, and upgrading. Subjects were instructed to choose multiple phases that they perceived as the most appropriate for that criterion.

The results from the first round were incorporated into the second round. The second-round survey investigated appropriateness of measures to their corresponding criterion. For the second round, 198 measures were identified while most criteria had multiple measures. Using seven point scale, this study tries to examine the appropriateness of measurements to assess each evaluation criterion. Also, the subjects were given the opportunity to modify the measures that were provided or to suggest new measures. This paper focuses on the findings of the first round in-depth survey for the proposed research question and associated hypothesis.

Fig. 1 shows an example of the first round survey.

Data analysis

Since most of the data collected through the first-round survey contained numerical ratings, quantitative analysis was applied. First, descriptive analysis was applied such as average and standard deviation to identify the importance of evaluation criteria. Based on average ratings, the evaluation criteria were ranked from the most important to the least. In addition, ratings of DL evaluation criteria of stakeholders were compared in order to better understand different perspectives and needs of stakeholders. To compare rating data of the importance of DL evaluation criteria, statistical test ANOVA was applied to examine the similarities and differences among the three groups of stakeholders. ANOVA is a method to examine if there is any significant mean difference among different groups. We measured the perceived degree of importance for each evaluation criterion from the three groups of participants except the criteria in Administration dimension. Then, ANOVA tests were applied to compare the means of responses from the three groups using 7-point scale and to confirm if any observed mean differences are statistically meaningful. For any statistically significant difference, a post-hoc test was performed to detect which pairs of groups are statistically significant in terms of their mean difference using the Tukey method. A post-hoc test compares all possible pairs of observed group means, so it can be used to find out exactly where significant differences lie (Field, 2013). However, due to the limited page space, the post-hoc results were not included in the results.

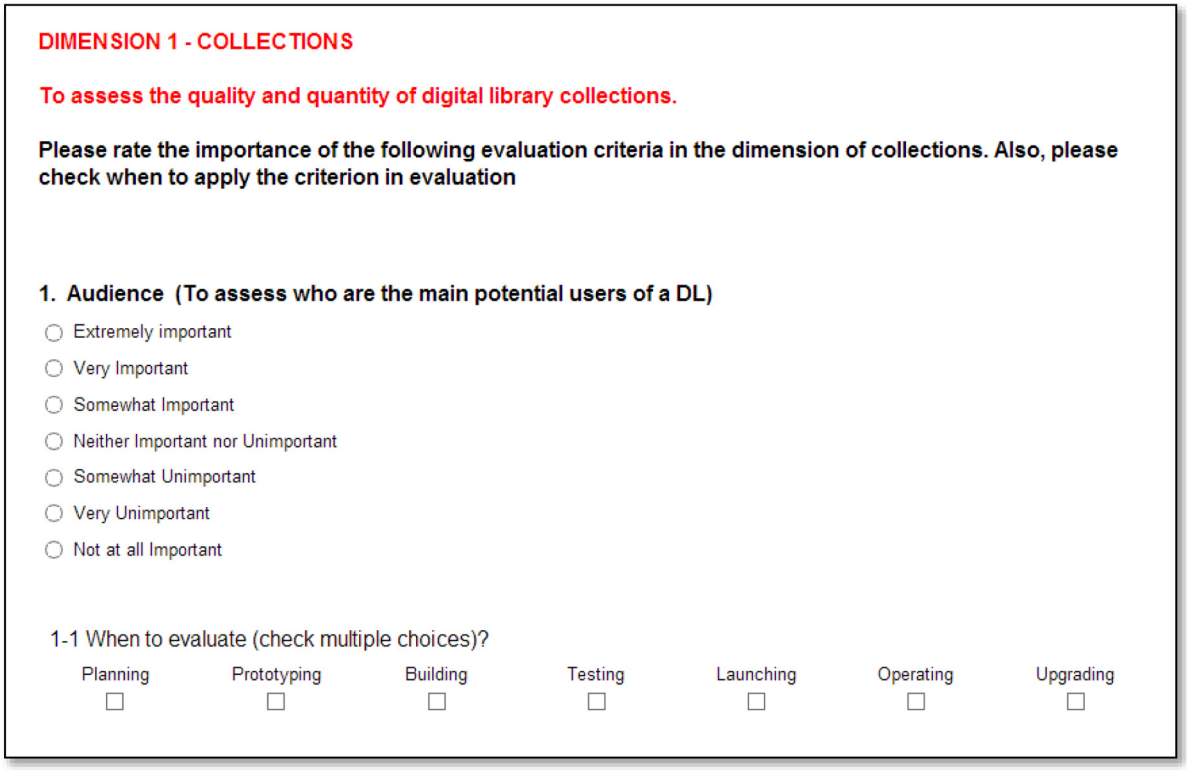

Fig. 1. An example question about the importance of DL evaluation criteria in the first round survey.

Instead, we reported all statistical significant cases in the narrative text. Only for the Administration dimension, *t*-test was used to compare the data from two groups because users do not have the expertise on this dimension. A t-test is a statistical analysis method to determine the difference between two means from two groups (Field, 2013).

Results

This result section presents similarities and differences on the ratings of the importance of evaluation criteria among three groups of DL stakeholders. In the tables below, the significant results ($p < 0.05$) are bolded for easy recognition.

Importance of DL evaluation criteria

(1) Dimension 1 - Collections

While all three groups of stakeholders perceived the use of standards and best practices as one of the most important criteria in evaluating DLs, they expressed different opinions on some of the evaluation criteria. The results showed that the three groups of stakeholders had significant differences in rating the importance of the following DL evaluation criteria associated with collections: "authority," "audience," "contextual information," "completeness," and "diversity." Ratings of importance showed different patterns by different groups. "Authority," which is related to the reliability of collection quality, was considered more important for the user group ($M = 6.53$) than the other two groups (scholars: $M = 6.33$, librarians: $M = 5.87$, $F(2, 88) = 3.717$, $p < .05$). At the same time, the scholar group perceived "audience" as more important than other two groups ($M = 6.47$) than the other two groups (librarians: $M = 5.65$, users: $M = 6.07$, $F(2, 88) = 5.12$, $p < .05$). Interestingly, the user group cared about "contextual information" ($M = 6.10$), "completeness" ($M = 5.77$) and "diversity" ($M = 5.77$) of collections much more than the scholar group ("contextual information": $M = 5.43$, $F(2, 88) = 8.21$; $p < .01$; "completeness": $M = 5.28$, $F(2, 88) = 4.21$, $p < .05$; "diversity": $M = 5.17$, $F(2, 88) = 8.98$; $p < .01$) and librarians ("contextual information": $M = 5.00$, $F(2, 88) = 8.21$, $p < .01$; "completeness": $M = 4.84$, $F(2, 88) = 4.21$; $p < .05$; "diversity": $M = 4.58$, $F(2, 88) = 8.98$, $p < .01$). However, the rating of the librarian group was relatively lower in both "authority" and "audience." Table 4 presents the importance of evaluation criteria in the dimension of collections.

(2) Dimension 2 - Information organization

For the dimension of information organization, the scholar group rated "appropriateness" the most important, while the librarian group chose "accessibility to metadata". The user group perceived accurate, consistent, and appropriate metadata important in regard to information organization in DLs. There were significant differences of ratings among the three groups for the criteria of "appropriateness" and "depth of metadata." As to "appropriateness," the rating of the scholar group ($M = 6.47$) was significantly higher than those of the other groups (librarians: $M = 6.13$; users: $M = 6.03$, $F(2, 88) = 3.110$, $p < .05$). For the criterion of "depth of metadata," the rating of the user group ($M = 6.21$) was significantly higher than those of the other two groups (scholars: $M = 5.73$, librarians: $M = 5.55$, $F(2, 88) = 4.79$, $p < .05$). Table 5 presents the importance of evaluation criteria in the dimension of information organization.

(3) Dimension 3 - Interface design

In terms of interface design, all three groups considered "search function" and "browsing function" important in evaluating DLs. "Navigation" and "intuitive operation" were also chosen as the important criteria across the three groups. "Visual appeal," "user control," and "personalized page" were rated least important in this dimension. However, in this dimension, perceptions of three groups were not significantly different except for the criterion of personalized page. The scholar group ($M = 4.93$) thought the criterion of personalized page was relatively more important compared with the other groups (librarians: $M = 3.68$, users: $M = 4.17$, $F(2, 88) = 6.58$, $p < .05$). Table 6 presents the importance of evaluation criteria in the dimension of interface design.

Table 4
Importance of evaluation criteria in the dimension of collections

Criteria	Scholars	Librarians	Users	Total	ANOVA
Digitization standards	6.10	6.48	6.20	6.26	$F(2, 88) = 2.007$; $p > .05$
Authority	**6.33**	**5.87**	**6.53**	**6.24**	**$F(2, 88) = 3.717$; $p < .05$**
Cost	6.13	6.23	6.10	6.15	$F(2, 88) = 0.172$; $p > .05$
Item quality	6.00	6.13	6.27	6.13	$F(2, 88) = 0.697$; $p > .05$
Format compatibility	6.07	6.06	6.10	6.08	$F(2, 88) = 0.014$; $p > .05$
Audience	**6.47**	**5.65**	**6.07**	**6.05**	**$F(2, 88) = 5.123$; $p < .01$**
Scope/coverage	5.72	5.32	5.83	5.62	$F(2, 88) = 2.415$; $p > .05$
Contextual information	**5.43**	**5.00**	**6.10**	**5.51**	**$F(2, 88) = 8.214$; $p < .01$**
Completeness	**5.28**	**4.84**	**5.77**	**5.29**	**$F(2, 88) = 4.208$; $p < .05$**
Diversity	**5.17**	**4.58**	**5.77**	**5.16**	**$F(2, 88) = 8.980$; $p < .01$**
Size	**5.00**	**4.74**	**5.57**	**5.10**	**$F(2, 88) = 3.622$; $p < .05$**

Table 5
Importance of evaluation criteria in the dimension of information organization.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Appropriateness	**6.47**	**6.13**	**6.03**	**6.21**	**$F(2, 88) = 3.110$; $p < .05$**
Accessibility to metadata	6.23	6.26	6.07	6.19	$F(2, 88) = 0.395$; $p > .05$
Metadata accuracy	6.23	5.97	6.28	6.16	$F(2, 88) = 1.562$; $p > .05$
Metadata standards	6.13	6.13	5.86	6.04	$F(2, 88) = 0.952$; $p > .05$
Consistency	5.83	5.73	6.24	5.93	$F(2, 88) = 2.262$; $p > .05$
Comprehensiveness	5.93	5.61	6.10	5.88	$F(2, 88) = 2.338$; $p > .05$
Depth of metadata	**5.73**	**5.55**	**6.21**	**5.83**	**$F(2, 88) = 4.789$; $p < .05$**
Metadata Interoperability	5.75	5.77	5.48	5.67	$F(2, 88) = 0.725$; $p > .05$
Controlled vocabulary	5.70	5.29	5.69	5.56	$F(2, 88) = 1.915$; $p > .05$

Table 6
Importance of evaluation criteria in the dimension of interface design

Criteria	Scholars	Librarians	Users	Total	ANOVA
Search function	6.60	6.55	6.48	6.54	F(2, 88) = 0.284; p > .05
Browsing function	6.30	6.20	6.53	6.34	F(2, 88) = 1.836; p > .05
Navigation	6.22	6.19	6.36	6.26	F(2, 88) = 0.463; p > .05
Intuitive operation	6.33	6.16	6.25	6.25	F(2, 88) = 0.342; p > .05
Search results presentation	6.27	6.16	6.10	6.18	F(2, 88) = 0.330; p > .05
Consistency	6.23	6.00	6.14	6.12	F(2, 88) = 0.675; p > .05
Reliability	6.17	5.90	6.28	6.11	F(2, 88) = 1.411; p > .05
Help function	5.59	5.42	5.93	5.64	F(2, 88) = 2.058; p > .05
Visual appeal	5.47	5.77	5.59	5.61	F(2, 88) = 0.899; p > .05
User control	5.10	4.61	5.14	4.95	F(2, 88) = 1.450; p > .05
Personalized page	**4.93**	**3.68**	**4.17**	**4.25**	**F(2, 88) = 6.582; p < .01**

Table 7
Importance of evaluation criteria in the dimension of system and technology.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Retrieval effectiveness	6.41	6.26	6.25	6.31	F(2, 88) = 0.469; p > .05
Reliability	6.17	6.13	6.25	6.18	F(2, 88) = 0.182; p > .05
Server performance	6.20	6.39	5.93	6.17	F(2, 88) = 1.825; p > .05
Response time	6.17	5.97	6.26	6.13	F(2, 88) = 1.100; p > .05
Fit-to-task	6.23	5.97	5.93	6.04	F(2, 88) = 1.451; p > .05
Connectivity	5.90	6.03	6.04	5.99	F(2, 88) = 0.281; p > .05
Page loading speed	6.07	6.00	5.86	5.97	F(2, 88) = 0.413; p > .05
Integrated search	6.17	5.94	5.75	5.95	F(2, 88) = 1.762; p > .05
Error rate/error correction	6.03	5.84	5.93	5.93	F(2, 88) = 0.436; p > .05
Flexibility	5.52	5.63	5.82	5.66	F(2, 88) = 1.105; p > .05
Linkage with other DLs	5.17	5.35	5.36	5.29	F(2, 88) = 1.762; p > .05

Table 8
Importance of evaluation criteria in the dimension of effects on users.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Research productivity	5.20	5.29	5.89	5.46	F(2, 88) = 3.084; p > .05
Learning effects	5.34	5.10	5.46	5.30	F(2, 88) = 0.814; p > .05
Knowledge change	4.93	4.94	5.26	5.04	F(2, 88) = 0.760; p > .05
Instructional efficiency	**4.63**	**4.77**	**5.32**	**4.91**	**F(2, 88) = 3.223; p < .05**
Perception of digital libraries	4.77	4.65	5.11	4.84	F(2, 88) = 0.884; p > .05
Information literacy/skill change	**4.80**	**4.13**	**5.00**	**4.64**	**F(2, 88) = 3.835; p < .05**

(4) Dimension 4 - System and technology

In the dimension of system and technology, “retrieval effectiveness,” “reliability,” and “server performance” turned out to be most important in DL evaluation. As DLs represent one type of the information retrieval systems, the participants perceived retrieval effectiveness, such as precision and recall, important in evaluating DLs. Reliability and server performance were ranked at second and third respectively in this dimension, which are necessary to provide stable services in DLs. Comparatively less important criteria were “error rate/error correction,” “flexibility,” and “linkage with other DLs.” Interestingly, no significant discrepancy was observed in the ratings among the three groups in this dimension. Table 7 presents the importance of evaluation criteria in the dimension of system and technology.

(5) Dimension 5 - Effects on users

In regards to the dimension of effects on users, “research productivity,” and “learning effects” were chosen as the most important criteria by all groups of the subjects. On the contrary, “perceptions of digital libraries” and “information literacy/skill change” were regarded relatively less important. There were significant differences in the ratings of “instructional efficiency” and “information literacy/skill change.” The user group (M = 5.32) rated “instructional efficiency” more important than the other groups did (scholars: M = 4.63, librarians: M = 4.77, F(2, 88) = 3.22, p < .05). The librarian group (M = 4.13) rated “information literacy/skill change” significantly lower compared with the other groups (scholars: M = 4.8, users = 5, F(2, 88) = 3.84, p < .05). Table 8 presents the importance of evaluation criteria in the dimension of effects on users.

(6) Dimension 6 - Services

In the dimension of services, the participants chose “service quality,” “usefulness,” and “user satisfaction” as three most important criteria. The evaluation criterion of services for users with disabilities was rated high ranked at fourth. On the other hand, “user education,” “types of unique services,” and “customized services” were ranked among least important criteria. There were several criteria that were rated significantly different by three groups in this dimension. For example, the scholar group (M = 5.21) and the user group (M = 4.89) identified “customized services” as more important than the librarian group (M = 4.16, F(2, 88) = 5.16, p < .01). Similarly, the scholar

Table 9
Importance of evaluation criteria in the dimension of services.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Service quality	6.41	5.97	6.36	6.24	F(2, 88) = 2.394; p > .05
Usefulness	6.41	6.00	6.29	6.23	F(2, 88) = 2.291; p > .05
User satisfaction	**6.45**	**5.77**	**6.32**	**6.18**	**F(2, 88) = 5.518; p < .01**
Types of services for users w/ disabilities	**6.00**	**5.94**	**6.43**	**6.12**	**F(2, 88) = 4.090; p < .05**
Reliability	**6.03**	**5.61**	**6.39**	**6.01**	**F(2, 88) = 4.600; p < .05**
Responsiveness	5.93	5.77	6.21	5.97	F(2, 88) = 2.565; p > .05
Timeliness	**6.03**	**5.68**	**6.11**	**5.94**	**F(2, 88) = 3.212; p < .05**
Types of services	**5.55**	**5.48**	**5.82**	**5.62**	**F(2, 88) = 4.090; p < .05**
Availability of DL staff	**5.17**	**5.77**	**5.89**	**5.61**	**F(2, 88) = 4.924; p < .01**
Confidence	5.89	5.32	5.43	5.55	F(2, 88) = 2.141; p > .05
Follow-up services	5.03	5.17	5.43	5.21	F(2, 88) = 0.932; p > .05
FAQ/Q&A	5.00	4.94	5.46	5.13	F(2, 88) = 2.112; p > .05
User education	5.03	4.90	5.36	5.10	F(2, 88) = 1.673; p > .05
Types of unique services	4.79	4.45	5.14	4.79	F(2, 88) = 2.548; p > .05
Customized services	**5.21**	**4.16**	**4.89**	**4.75**	**F(2, 88) = 7.156; p < .01**

Table 10
Importance of evaluation criteria in the dimension of preservation.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Completeness	6.07	6.23	6.36	6.22	F(2, 88) = 0.746; p > .05
Ability to migrate	**6.21**	**6.45**	**5.89**	**6.19**	**F(2, 88) = 3.229; p < .05**
Preservation policy	5.97	6.00	6.04	6.00	F(2, 88) = 0.050; p > .05
Preservation infrastructure	5.66	6.16	5.71	5.85	F(2, 88) = 2.187; p > .05
Institutional support	5.72	5.74	5.82	5.76	F(2, 88) = 0.066; p > .05
Types of archiving methods	5.66	5.90	5.64	5.73	F(2, 88) = 0.730; p > .05
Cost per record	**5.69**	**4.87**	**5.79**	**5.44**	**F(2, 88) = 5.479; p < .01**

Table 11
Importance of evaluation criteria in the dimension of administration.

Criteria	Scholars	Librarians	Total	T-Test
Budget	6.21	6.13	6.18	t(59) = 0.323; p > .05
Planning	6.11	6.10	6.10	t(59) = 0.064; p > .05
Staffing	6.11	5.94	6.02	t(59) = 0.808; p > .05
Staff training	5.72	5.87	5.79	t(59) = −0.703; p > .05
Marketing	5.66	5.68	5.67	t(59) = −0.091; p > .05
Regular assessment	**5.97**	**5.23**	**5.59**	**t(59) = 2.360; p < .05**
Management Policy	5.62	5.19	5.40	t(59) = 1.324; p > .05
Fundraising/sponsor	5.34	5.23	5.28	t(59) = 0.372; p > .05
Incentive	**4.39**	**3.97**	**4.32**	**t(59) = 2.403; p < .05**

group (M = 6.45) and the user group (M = 6.32) selected "user satisfaction" as more important than the librarian group (M = 5.77, F(2, 88) = 5.518, p < .01). On the contrary, the scholar group (M = 5.17) perceived "availability of DL stuff" less important than the other two groups (users: M = 5.89, librarians: M = 7.77, F(2, 28) = 4.92, p < .01). Table 9 presents the importance of evaluation criteria in the dimension of services.

(7) Dimension 7 – Preservation

In the dimension of preservation, "completeness," "ability to migrate," and "preservation policy" were ranked 1st, 2nd, and 3rd respectively. "Institutional support," "types of archiving methods," and "cost per record" were perceived less important by all three groups according to the survey. Interestingly, the librarian group rated the criterion of "ability to migrate" the highest, but they perceived "cost per record" less important compared to the other groups. On the one hand, the librarian group (M = 6.45) considered "ability to migrate" more important than the scholar group (M = 6.21) and the user group (M = 5.89, F(2, 88) = 3.23, p < .05). On the contrary, the librarian group (M = 4.87) perceived "cost per record" less important than the scholar group (M = 5.69) and the user group (M = 5.79, F(2, 88) = 5.48, p < .01). Overall, librarians and users exhibited higher scores in the dimension of preservation as shown in Table 10.

(8) Dimension 8 - Administration

Since users typically do not have sufficient knowledge in administration of DLs, they were excluded from the survey in this dimension. The librarian and scholar groups rated "budget," "planning," and "staffing" as the three most important criteria in the dimension of administration. There was no significant difference for these top three criteria between the scholar and librarian groups. On the contrary, "management policy," "fundraising/sponsor," and "incentive" were considered less important by the two groups. There were significant differences in "regular assessment" and "incentive." The *t*-test results showed that there was a significant difference between the scholars (M = 5.97) and the librarians (M = 5.23) in their rating of "regular assessment," t(59) = 2.360, $p < .05$. Also, a significant difference was found in the rating of "incentive" between the two groups, t (59) = 2.403, p < .05. Table 11 presents the importance of evaluation criteria in the dimension of administration.

(9) Dimension 9 - User engagement

Table 12
Importance of evaluation criteria in the dimension of user engagement.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Resource use	6.31	6.00	5.81	6.04	F(2, 88) = 2.425; p > .05
User feedback	6.03	5.97	5.89	5.97	F(2, 88) = 0.183; p > .05
Site visit	5.81	5.90	5.50	5.74	F(2, 88) = 1.878; p > .05
Integration with external applications	5.62	5.42	5.50	5.51	F(2, 88) = 0.261; p > .05
Help feature use	**5.34**	**5.16**	**5.79**	**5.43**	**F(2, 88) = 3.551; p < .05**
User participation channels	5.41	5.37	5.39	5.39	F(2, 88) = 0.013; p > .05
User knowledge contribution	5.29	5.00	5.00	5.26	F(2, 88) = 1.527; p > .05
E-commerce support	4.71	4.23	4.89	4.61	F(2, 88) = 1.961; p > .05

Table 13
Importance of evaluation criteria in the dimension of context.

Criteria	Scholars	Librarians	Users	Total	ANOVA
Copyright	6.24	6.29	6.25	6.26	F(2, 88) = 0.27; p > .05
Information ethics compliance	**6.04**	**6.06**	**6.61**	**6.23**	**F(2, 88) = 4.328; p < .05**
Organizational mission	6.03	5.90	5.43	5.79	F(2, 88) = 2.905; p > .05
Targeted user community	**6.17**	**5.39**	**5.75**	**5.77**	**F(2, 88) = 5.260; p < .01**
Content sharing	5.52	5.57	6.00	5.69	F(2, 88) = 2.319; p > .05
Collaboration	**4.93**	**5.23**	**5.75**	**5.30**	**F(2, 88) = 5.238; p < .01**
Social impact	5.10	4.65	5.18	4.97	F(2, 88) = 1.705; p > .05

In the dimension of user engagement, “resource use,” “user feedback,” and “site visit” were the three highly rated criteria by the three groups. All three groups also perceived “user feedback” important in DL evaluation. On the other hand, “user participant channels,” “user knowledge contribution,” and “e-commerce support” were rated relatively less important. There was a significant difference in the ratings of “help feature uses” among the three groups (users: M = 5.79, scholars: M = 5.34, librarians M = 5.16, F(2, 88) = 3.55, p < .05). Table 12 presents the importance of evaluation criteria in the dimension of user engagement.

(10) Dimension 10 - Context

Finally, subjects selected “copyright,” “information ethics compliance,” and “organizational mission” as the most three important criteria in this dimension. In particular, the user group (M = 6.61) rated relatively higher for “information ethics compliance” than the scholar group (M = 6.04) and the librarian group (M = 6.06, F(2, 88) = 4.33, p < .05). The scholar group (M = 6.17) perceived “target user community” more important compared to the other groups (librarians: M = 5.39, users: M = 5.75, F(2, 88) = 5.26, p < .01). On the other hand, “content sharing,” “collaboration,” and “social impact” were considered less important. Among them, the user group (M = 5.75) rated the importance of “collaboration” more than the scholar group (M = 4.93) and the librarian group (M = 5.23, F(2, 88) = 5.24, p < .01). Overall, the three groups' ratings in this dimension are very diverse, and none of the group shows dominate higher ratings across criteria. Table 13 presents the importance of evaluation criteria in the dimension of context.

Discussion

Identifying evaluation criteria is essential for the development of comprehensive frameworks and conducting holistic evaluation of DLs. This study examined the perceived importance of evaluation criteria within ten dimensions from the perspective of three groups of stakeholders. In grouping the criteria along the dimensions, it built on prior research of Saracevic (2004) and Zhang (2010) but also expanded it by analyzing evaluation criteria in four new dimensions, including preservation. As discussed in the Literature Review section, there is a growing interest in digital preservation as an area of research and practice. Interestingly, preservation emerges as a unique dimension in this study, perceived by librarians as an important aspect of DL evaluation.

This study demonstrates noticeable differences in stakeholders' perceptions about the importance of evaluation criteria and confirms the divergence of opinions that was found by Zhang (2010). The findings do not support the hypothesis (H1) generated for this study stating that there is no significant difference in rating of evaluation criteria among the three groups of stakeholders. The patterns of ranking are similar between scholars and users, who assigned higher rankings than librarians. Generally, librarians ranked the evaluation criteria lower across dimensions except for Preservation. The differences in assigning ranks are consistent and systematic, and seem to be related to the stakeholders' different needs, interests, and familiarity with DL concepts. Users, for example, ranked higher criteria related directly to use of collections, such as authority in collections, or metadata accuracy in information organization, or the quality of services.

The overall lower ranking of evaluation criteria by librarians might be interpreted as a more realistic assessment of practitioners involved in the development of DLs. For example, librarians assigned a lower ranking to authority or metadata accuracy than scholars or users, which may come as a surprise. Metadata accuracy, although a desirable goal in information organization of digital collections, in practice might be very difficult to achieve because of limited descriptions of original collections, the lack of resources for creating in-depth metadata, and inter-indexer inconsistency. Practitioners are aware that creating good-quality and accurate metadata is extremely challenging (Matusiak et al., 2017; Miller, 2011). On the other hand, from the user and scholar perspective, metadata accuracy and quality is a very important criterion related to the successful resource discovery. This finding confirms previous research on the importance of metadata quality and its relevance to improving access to users, especially in large-scale DL environment (Zavalina, 2014).

Librarians involved in the development of DLs have a realistic view of the current challenges in the field and of what can be achieved in practice. They consistently ranked higher criteria related to practical aspects of DL development and management, such as preservation. A higher ranking of all criteria in the preservation indicates that librarians are keenly aware of the challenges of digital preservation and its importance to DL evaluation. Preservation as an evaluation dimension that does not have an immediate impact on current users but is

important to sustainability and long-term management of digital assets. Interestingly, librarians ranked criteria in preservation not only higher than users but also scholars. This is surprising since digital preservation is a current and important topic in research literature (Chowdhury, 2010; Ross, 2012; Xie & Matusiak, 2016). Librarians' attention to the practical aspects of DL evaluation is understandable. However, lower rankings assigned to criteria connected to effect on users or services might be an area of concern. The preoccupation with DL development and management should not overshadow the fact that DLs are created for users and should serve users' needs and meet their expectations.

Users ranked high the criteria within the dimensions that affect their interaction with DL systems, including collections, information organization, interface design, and services. This finding is consistent with prior research that found usability and collection quality as the most important aspects of the evaluation from an end-user perspective (Xie, 2008). Users ranked authority as the highest criterion within the Collections dimension, which indicates user expectations for DLs as sources of authoritative information, especially in contrast to the materials available on the open Web where authority and reliability are uncertain. Users in this study also ranked high “research productivity” and “learning effects” in the Dimension 5 – Effects on Users, which clearly reflects the goals of academic users recruited for the study.

DL scholars were generally closer to users than librarians in ranking the criteria across dimensions. DL scholars tended to rank higher criteria that focused on users, such as “audience,” “user satisfaction,” “customized services,” or “targeted user community.” In the dimension “User Engagement,” scholars ranked almost all criteria higher than the two other groups, which demonstrates an understanding of the importance of user participation and feedback. However, there were some notable exceptions. In the dimension “Information Organization,” scholars' ranking of criteria was similar to those of librarians, which can be explained by expert knowledge of both groups and better understanding of the importance of metadata standards and technical aspects of developing and sharing metadata. In fact, the criterion “metadata standards” has exactly the same score (6.13) between scholars and librarians. There was also no significant difference in ranking the criteria in the “Administration” dimension between the scholar and librarian groups except for “regular assessment,” which was ranked higher by scholars. As discussed in the literature review section, the evaluation studies of DLs have been primarily conducted by researchers (Albertson, 2015; Dobreva & Chowdhury, 2010; Xie, 2006, 2008; Zhang, 2010). This is an important area of research in DLs and hence higher ranking of assessment does not come as a surprise. Lower ranking of assessment by librarians may not reflect the lack of interest but rather limited resources to conduct evaluation studies and implement changes.

In addition to pointing to differences in ranking of evaluation criteria among the stakeholder groups, this study also finds some consensus in perceptions of criteria importance. The high ranking of “copyright” and “information ethics compliance” of all three groups indicates that the academic setting makes the awareness of these concepts popular among not only scholars but also librarians and users. Digitization standards criterion was rated high consistently across the three groups, demonstrating perhaps a growing maturity of the field and awareness that adherence to the standards is directly linked to the quality of digital objects. There is also inter-group agreement around the top criteria in user interface dimension, including search, browse, and navigation functions. The importance of these criteria is supported by prior research on search and information retrieval and practical evaluation studies (Behnert & Lewandowski, 2017; Matusiak, 2017).

This study also has some limitations in the research design. Since the limited number of scholars in the area of DL research and digital librarians in the library field, it was hard to recruit a sufficient number of participants. More than a hundred scholars domestically and internationally were contacted, but only thirty responded to our survey request. Also, the surveys were solely quantitative (e.g. 7-point scale); it was not possible to further investigate underlying reasons behind subjects' opinions. Therefore, the results are limited to numeration of the importance and appropriateness of criteria. Detailed qualitative data is missing, and accordingly, the result of this study is more descriptive and statistical rather than explaining detailed reasons in the selection of criteria. Furthermore, this study presents the DL dimensions and evaluation criteria identified in previous research literature. As the field of DL evaluation evolves, new methods of assessing value and use of DLs are emerging, such as altmetrics and reuse measures.

Conclusion

This study presents comprehensive and multifaceted evaluation of DL dimensions and criteria and examines the importance of the criteria from the perspective of scholars, librarians, and users in the academic setting. It expands the research on evaluation of DLs by introducing new dimensions, including preservation. It points to the importance of preservation in evaluating DLs from the perspective of librarians. It also highlights the criteria that are important to end users in academic libraries. This study finds consensus and divergence in perceptions of criteria importance among the three participant groups. The differences in ranking of criteria importance reveal a gap between user expectations and the world of DL practice as well as between what's desirable and what's possible in the current technological environment. Furthermore, this study indicates an inherent and perhaps unavoidable tension between different stakeholders of DLs. As Kelly (2014) emphasizes the goal of evaluation studies is to provide data for improving user experience and to justify the creation of digital collections to multiple stakeholders. While acknowledging the inherent differences, the evaluation studies can also contribute to narrowing the gap between the developers and users as well as the gap between scholars and developers of DLs. This study provides a comprehensive list of criteria to guide practical evaluation of DLs in the academic setting. Future research could expand the findings presented in this study by incorporating criteria related to alternative use and reuse of digital objects. In addition, qualitative data collection (e.g. interviews) and data analysis need to be added in relation to the reasons behind the selection of DL evaluation criteria.

References

Albertson, D. (2015). Synthesizing visual digital library research to formulate a user-centered evaluation framework. *New Library World, 116*(3/4), 122–135.

Beaudoin, J. E. (2012a). *Context and its role in the digital preservation of cultural objects. 18(11)*, D-Lib Magazine1. http://www.dlib.org/dlib/november12/beaudoin/11beaudoin1.html, Accessed date: 1 April 2018.

Beaudoin, J. E. (2012b). *A framework for contextual metadata used in the digital preservation of cultural objects. 18(11/12)*, D-Lib Magazine3. http://dlib.org/dlib/november12/beaudoin/11beaudoin1.print.html (1 April 2018).

Behnert, C., & Lewandowski, D. (2017). A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments. *Journal of Documentation, 73*(3), 509–527.

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

Bui, Y., & Park, J. R. (2006). An assessment of metadata quality: A case study of the national science digital library metadata repository. In H. Moukdad (Ed.). *CAIS/ACSI 2006 information science revisited: Approaches to innovation*.

Candela, L., Castelli, D., Pagano, P., Thanos, C., Ioannidis, Y., Koutrika, G., ... Schuldt, H. (2007). *Setting the foundations of digital libraries: The DELOS manifesto. 13(3/4)*D-Lib Magazinehttp://www.dlib.org/dlib/march07/castelli/03castelli.html (22 March 2018).

Carr, R. (2006). *What users want: An academic ‘hybrid’ library perspective. 46*Ariadne. Available from: http://www.ariadne.ac.uk/issue46/carr/ (22 March 2018) .

Chowdhury, G. (2010). From digital libraries to preservation research: The importance of users and context. *Journal of Documentation, 66*(2), 207–223.

Chowdhury, G. (2014). Sustainability of digital libraries: A conceptual model and a research framework. *International Journal on Digital Libraries, 14*(3–4), 181–195.

Dobreva, M., & Chowdhury, S. (2010). A user-centric evaluation of the Europeana digital library. In G. Chowdhury, C. Koo, & J. Hunter (Eds.). *The role of digital libraries in a time of global change* (pp. 148–157). Berlin: Springer.

Dressler, V. (2017). The state of affairs with digital preservation at ARL member libraries. *Digital Library Perspectives, 33*(2), 137–155.

Ferati, M., & Beyene, W. M. (2017). Developing heuristics for evaluating the accessibility

of digital library interfaces. In M. Antona, & C. Stephanidis (Eds.). *Universal access in human–computer interaction. Design and development approaches and methods* (pp. 171–181). Cham: Springer UAHCI 2017.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* London: Sage.
Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., ... Sølvberg, I. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries, 8*, 21–38.
Gkoumas, G., & Lazarinis, F. (2015). Evaluation and usage scenarios of open source digital library and collection management tools. *Program, 49*(3), 226–241.
Greenstein, D. (2000). Digital libraries and their challenges. *Library Trends, 49*(2), 290–303.
Inal, Y. (2018). University students' heuristic usability inspection of the National Library of Turkey website. *Aslib Journal of Information Management, 70*(1), 66–77.
Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri, 55*, 96–121.
Joo, S., & Lee, J. (2011). Measuring the usability of academic digital libraries: Instrument development and validation. *The Electronic Library, 29*(4), 523–537.
Joo, S., & Xie, I. (2013). Evaluation constructs and criteria for digital libraries: A document analysis. In C. Cool, & K. B. Ng (Eds.). *Recent developments in the design, construction and evaluation of digital libraries* (pp. 126–140). Hershey, PA: IGI Global.
Kani-Zabihi, E., Ghinea, G., & Chen, S. Y. (2006). Digital libraries: What do users want? *Online Information Review, 30*(4), 396–412.
Kelly, E. J. (2014). Assessment of digitized library and archives materials: A literature review. *Journal of Web Librarianship, 8*(4), 384–403.
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.
Lai, C. F., Chiu, P. S., Huang, Y. M., Chen, T. S., & Huang, T. C. (2014). An evaluation model for digital libraries' user interfaces using fuzzy AHP. *The Electronic Library, 32*(1), 83–95.
Lesk, M. (2005). *Understanding Digital Libraries* (2nd ed.). San Francisco: Morgan Kaufman.
Matusiak, K. K. (2017). User navigation in large-scale distributed digital libraries: The case of the digital public library of America. *Journal of Web Librarianship, 12*(3), 1–15.
Matusiak, K. K., Taylor, A., Newton, C., & Polepeddi, P. (2017). Finding access and digital preservation solutions for a digitized oral history project: A case study. *Digital Library Perspectives, 33*(2), 88–99.
Miller, A. (2017). A case study in institutional repository content curation. *Digital Library Perspectives, 33*(1), 63–76.
Miller, S. J. (2011). *Metadata for digital collections: A how-to-do-it manual.* New York: Neal-Schuman Publishers.
Mune, C., & Agee, A. (2016). Are e-books for everyone? An evaluation of academic e-book platforms' accessibility features. *Journal of Electronic Resources Librarianship, 28*(3), 172–182.
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. (1995). 10 usability heuristics for user interface design. http://www.nngroup.com/articles/ten-usability-heuristics/, Accessed date: 5 April 2018.
Noh, Y. (2010). A study on developing evaluation criteria for electronic resources in evaluation indicators of libraries. *Journal of Academic Librarianship, 36*(1), 41–52.
Noonan, D. (2014). Digital preservation policy framework: A case study. *Educause Review, 49*(4)http://er.educause.edu/articles/2014/7/digital-preservation-policy-framework-a-case-study, Accessed date: 2 April 2018.
Petras, V., & Stiller, J. (2017). *A decade of evaluating Europeana-constructs, contexts, methods & criteria* (pp. 233–245). Cham: Springer.
Recker, M. (2006). Perspectives on Teachers as Digital Library Users. *D-Lib Magazine, 12*(9), (Accessed 1 April 2018) http://mirror.dlib.org/dlib/september06/recker/09recker.html.
Recker, M., Walker, A., Giersch, S., Mao, X., Halioris, S., Palmer, B., & Robertshaw, M. B. (2007). A study of teachers' use of online learning resources to design classroom activities. *New Review of Hypermedia and Multimedia, 13*(2), 117–134.
Rinehart, A. K., Prud'homme, P.-A., & Huot, A. R. (2014). Overwhelmed to action: Digital preservation challenges at the under-resourced institution. *OCLC Systems & Services, 30*(1), 28–42.
Ross, S. (2012). Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking, 17*(1), 43–68.
Saracevic, T. (2004). Evaluation of digital libraries: An overview. *Notes of the DELOS WP7 workshop on the evaluation of digital libraries, Padua, Italy*http://www.scils.rutgers.edu/~tefko/DL_evaluation_Delos.pdf, Accessed date: 20 March 2018.
Scrivin, M. (1991). *Evaluation thesaurus* (4th edition). Newbury Park: Sage Publications.
Weiss, A. (2014). *Using massive digital libraries: a LITA guide.* Chicago: American Library Association.
Xie, I. (2006). Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library & Information Science Research, 28*(3), 433–452.
Xie, I. (2008). Users' evaluation of digital libraries: Their uses, their criteria, and their assessment. *Information Processing & Management, 44*(3), 1346–1373.
Xie, I., Babu, R., Joo, S., & Fuller, P. (2015). Using digital libraries non-visually: Understanding the help-seeking situations of blind users. *Information Research: An International Electronic Journal, 20*(2)http://www.informationr.net/ir/20-2/paper673.html, Accessed date: 2 April 2018.
Xie, I., & Matusiak, K. (2016). *Discover digital libraries: Theory and practice.* Amsterdam: Elsevier.
Xu, B., & Recker, M. (2012). Teaching analytics: A clustering and triangulation study of digital library user data. *Journal of Educational Technology & Society, 15*(3), 103.
Zavalina, O. L. (2014). Complementarity in subject metadata in large-scale digital libraries: A comparative analysis. *Cataloging & Classification Quarterly, 52*(1), 77–89.
Zhang, Y. (2010). Developing a holistic model for digital library evaluation. *Journal of the American Society for Information Science and Technology, 61*(1), 88–110.