User–Centered Evaluation of Digital Libraries: a Literature Review

Draft version of the paper published in Journal of Information Science, doi:10.1177/0165551512438359

Ruben Heradio *[1], David Fernandez-Amoros†[1], Javier Cabrerizo‡[1], and Enrique Herrera–Viedma[2]

[1]Department of Software Engineering and Computer Systems, Distance Learning University of Spain (UNED), Spain
[2]Department of Computer Science and Artificial Intelligence, University of Granada, Spain

Abstract

In the past two decades, the use of Digital Libraries (DLs) has grown significantly. Accordingly, questions about utility, usability, and cost of DLs have started to arise and greater attention is given to the evaluation of this type of information system. Since DLs are destined to serve user communities, one of the main aspects to be considered in DL evaluation is users' opinion. The literature on this topic has produced a set of different criteria to judge DLs from the users' perspective; measuring instruments to elicit users' opinion and approaches to analyze the elicited data to conclude an evaluation. This paper provides a comprehensive literature review on the user–centered evaluation of DLs. We believe its main contribution is to bring together previously-disparate streams of work to help shed light on this thriving area. In addition, the paper discusses the different studies and proposes some challenges to be faced in the future.

1 Introduction

Digital Libraries (DLs) can be defined as collections of information that have associated services delivered to user communities using a variety of technologies [45]. In general, DLs are the logical extension of physical libraries in an electronic information society. Such extensions offer new levels of access to broader audiences of users [35].

The use of DLs has grown significantly in the past two decades [48]. By the end of the 1980's, DLs were barely a part of the landscape of librarianship, information science, or computer science. A decade later, by the end of the 1990's, research, practical developments, and general interest in DLs exploded globally. The accelerated growth of numerous and highly varied efforts related to DLs has continued unabated in the 2000's [41].

Once the importance and applicability of this type of information system has been definitely established, questions about utility, usability, and cost of DLs have started to arise and greater attention is given to their evaluation. To define what makes a DL a good quality system can be difficult and hard to summarize, since it depends on which of the many aspects of a DL are being considered [37]. This has lead to the expansion of DL evaluation to sectors like database structure, network architecture, protocols interoperability, development of intelligent and adaptive technologies, performance of retrieval algorithms,

*rheradio@issi.uned.es
†david@lsi.uned.es
‡cabrerizo@issi.uned.es

collection development, digitization policy assessment, usability, information architecture, interaction design, information behavior and many other [49].

The final aim of a DL system is enabling people to access human knowledge any time and anywhere, in a friendly multi–modal way, by overcoming barriers of distance, language and culture, and by using multiple network connected devices [14]. DLs are destined to serve users: if unused, these systems fall into oblivion and terminate their operation [6]. Therefore, one of the main aspects to be considered in DL evaluation is users' perspective, determining the extent to which the DL addresses the real needs of its users.

User–centered evaluation of DLs has drawn considerable attention during the last years [54]. Research on this area has produced a set of different criteria to judge DLs from the users' perspective; measuring instruments to elicit users' opinion and approaches to analyze the elicited data to conclude an evaluation. This paper provides a comprehensive literature review on such issues.

According to Saracevic [42], the literature on DL evaluation can be divided in two distinct types: (i) *meta* or "about" literature (i.e., works that suggest evaluation concepts, models, approaches, methodologies or discuss evaluation) and (ii) *object* or "on" literature (i.e., works that report on actual evaluation and contain data). This paper reviews *meta*–literature on user-centered evaluation of DLs.

An effective review creates a firm foundation for advancing knowledge. It facilitates theory development, closes areas where a plethora of research exists, and uncovers areas where research is needed. In order to fulfill such goals, our review follows a rigorous and auditable methodology proposed by Kitchenham [31] and Webster et al. [51]. Specifically, this paper addresses the following research questions: (1) What criteria are proposed to evaluate DLs in an user-centered fashion?, (2) How are those criteria measured and processed?, and (3) What are the most important challenges to be faced in the future?

The remainder of the paper is structured as follows: Section 2 presents the systematic method we have used to review the literature. Section 3 summarizes what criteria are proposed for the user-centered evaluation of DLs, the importance of each criterion and the inter–criteria correlation. Section 4 surveys the quantitative and qualitative measures that are derived from those criteria, the measuring instruments to elicit users' opinion and how the measurements are combined to conclude a DL evaluation. Section 5 discusses the results of the review and describes fundamental challenges to be faced in the future. Finally, section 6 presents the conclusions of the paper.

2 Review method

To perform our review we have followed a systematic and structured method inspired by the guidelines of Kitchenham [31] and Webster et al. [51]. Below, we detail the main data regarding the review process and its structure.

2.1 Research questions

The aim of this review is to answer the following Research Questions (RQs):

- *RQ1: What criteria are used to evaluate DLs in an user-centered fashion?* This question motivates the following sub-questions:
  - Have all criteria the same importance on the evaluation?
  - Is there any correlation among the criteria?
- *RQ2: How are those criteria measured?* This question motivates the following sub-questions:
  - Are the measures quantitative or qualitative?
  - What instruments are used to elicit users' opinion?

– How are the measurements analyzed?

After reviewing all this information, we also want to answer a more general question:

- *RQ3: What are the challenges to be faced in the future?*

Sections 3, 4 and 5 attempt to answer questions RQ1, RQ2 and RQ3, respectively.

2.2 Source material

As recommended by Webster et al. [51], we have used both manual and automated methods to make a selection of candidate papers in leading journals, conferences and other related events. We reviewed 67 papers: 32 were discarded, resulting in a total of 35 papers that were in the scope of this review. These 35 papers are referred as *primary studies* [31].

Table 1 and Figure 1 classify primary studies according to the year and type of publication. Of the 35 papers included in the review, 25 were published in journals, 3 in conferences, 5 in workshops, 1 in a book and 1 as a technical report.

Year	Journals	Conferences	Workshops	Others
1999	[30]			
2000	[26]			
2001	[34, 29]	[13, 21]		
2002	[48]		[11, 43]	
2003				[6]
2004	[9]	[5]	[4, 42, 49]	
2005	[23, 24, 40]	[46]		
2006	[52, 27]			
2007	[14]			[8]
2008	[1, 3, 20, 25, 50, 53, 55]			
2009	[12, 32, 37]			
2010	[7, 15, 54]			

Table 1: Classification of papers per year and type of publication

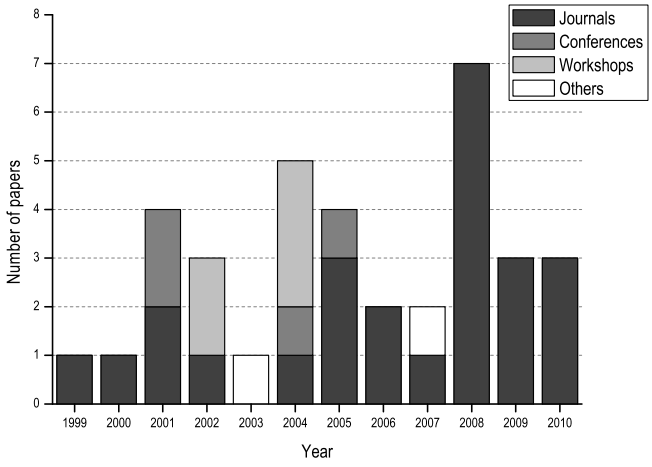


Figure 1: Number of papers per year and type of publication

2.3 Inclusion and exclusion rationale

The aim of this paper is to review the user–centered evaluation of DLs. Nevertheless, the term *user* has different meanings in the DL context. For instance, the DELOS Digital Library Reference Model [8] identifies the following types of actors that interact with DLs:

1. *DL end–users* exploit the DL functionality for the purpose of providing, consuming and managing the DL content and some of its other constituents. DL end-users may be further divided into:
   (a) *Content consumers* are the purchasers of the DL content.
   (b) *Content creators* are the producers of the DL content; they feed it with the resources, mainly information objects, to which other users of the DL will have access.
   (c) *Librarians* are end-users in charge of curating the DL content. In fact, these actors have to curate all the resources forming the DL, e.g. establish the policies.
2. *DL designers* exploit their knowledge of the application semantic domain in order to define, customize and maintain the DL so that it is aligned with the information and functional needs of its potential DL end-users.
3. *DL system administrators* select the software components needed to construct the DL system. Their choice of elements reflects the expectations that DL end-users and DL designers have for the DL, as well as the requirements the available resources impose on the definition of the DL.
4. *DL application developers* develop the software components that will be used as constituents of the DL systems, to ensure that the appropriate levels and types of functionality are available.

According to the user classification proposed by DELOS, we restrict our survey to the DL evaluation from the *content consumers* point of view.

We have included articles on the following topics, published between January 1st 1999 and December 31st 2010:

- User–centered proposals for DL evaluation.
- Holistic approaches for DL evaluation.
- DL usability/usefulness measurement and analysis.
- DL usability/usefulness sub–criteria correlation.
- DL usability/usefulness sub–criteria importance.

3 Criteria for DL evaluation

This section outlines the criteria that have been proposed to evaluate DLs from the users' perspective. Firstly, to provide a classification of such criteria, a conceptual model for DL evaluation, that is backed up by the Working Group on Evaluation of the DELOS Network of Excellence, is presented. Then, the criteria are summarized.

Fuhr et al. [13] propose a generic conceptual model for DL evaluation, which is composed of three non-orthogonal components: the *users*, the DL *content* and the technological *system* that supports the DL content. According to Fuhr's model, "content is king" and consequently, the nature, extent and form of the DL content predetermine both the range of potential users and the required technology. Although Fuhr's model addresses system-centered evaluations [14], it has remarkably influenced several user-centered proposals. In particular, Tsakonas et al. [49] propose a user-centered model focused on the relations between the components of Fuhr's model. These relations are shown in Figure 2:

- The *content-system* pair is related to performance criteria (precision, recall, response time...).
- The *user-system* pair is related to the *usability*[1] criterion, which defines the quality of the interaction between the *user* and the *system*. Usability evaluates whether the system is manipulated effectively by the user, in an efficient and enjoyable way which supports exploiting all the available functionalities. A usable system is easy to learn, flexible and adapts to user preferences and skills.
- The *user-content* pair is related to the *usefulness* criterion, which evaluates the relevance of the DL content to the user tasks and needs.

This paper is focused on the *user-system* and the *user-content* interactions (see the shadowed area in Figure 2). At the moment, there is no consensus on the definition of the usability and usefulness criteria, nor on their importance or correlation. The following subsections review available proposals on such issues.

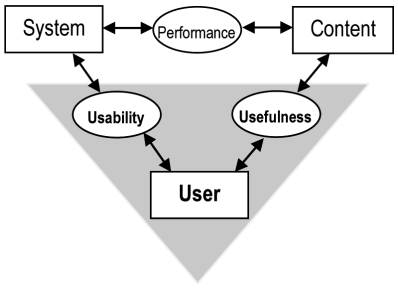


Figure 2: DL evaluation model proposed by Tsakonas et al. [49]

3.1 Usability

According to Shackel [44], the definition of informatics usability was probably first attempted in 1971 by Miller [36] in terms of measures for "ease of use". Since then, a wide variety of definitions for informatics usability has been proposed (e.g., Jeng [24] reviews 15 different definitions for informatics usability). In this paper, we restrict our usability review to the DL context.

Most authors consider usability as a complex concept composed of a number of criteria. Table 2 summarizes the usability criteria and sub-criteria considered by Evans et al. [11], Jeng [23, 24], Saracevic [42], Snead et al. [46], Tsakonas et al. [49, 50] and Xie [52].

1. **Evans et al. [11]** proposes a usability evaluation framework adapted from the heuristic approach of Nielsen [38]. It takes into account the following criteria:
   (a) *Visibility of system status.* The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
   (b) *Match between system and the real world.* The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms.
   (c) *User control and freedom.* Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue.

[1] In the literature, *ease of use* and *utility* are sometimes used as synonyms for *usability* and *usefulness*, respectively [11].

(d) *Consistency and standards*. Users should not have to wonder whether different words, situations, or actions mean the same thing.

(e) *Error prevention*. Even better than good error messages is a careful design that prevents a problem from occurring in the first place.

(f) *Recognition rather than recall*. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

(g) *Flexibility and efficiency of use*. Accelerators – unseen by the novice user – may often speed up the interaction for the expert user so that the system can cater to both inexperienced and experienced users.

(h) *Aesthetic and minimalist design*. Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

(i) *Help users recognize, diagnose, and recover from errors*. Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

(j) *Help and documentation*. Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the users' task, list concrete steps to be carried out, and not be too large.

2. **Jeng [23, 24]** proposes an evaluation model that applies the usability definition of ISO 9241-11 [39]. It examines the following criteria:

   (a) *Effectiveness*. It evaluates if the system can provide information and functionality effectively.

   (b) *Efficiency*. It evaluates if the system can be used to retrieve information efficiently.

   (c) *Satisfaction*. It encompasses the following sub-criteria:

      i. *Ease of use*. It evaluates users' perception on the ease of use of the system.

      ii. *Organization of information*. It evaluates if the system's structure, layout, and organization meets the users' satisfaction.

      iii. *Labeling*. It evaluates from users' perception if the system provides clear labeling and if terminology used is easy to understand.

      iv. *Visual appearance*. It evaluates the site's design to see if it is visually attractive.

      v. *Contents*. It evaluates the authority and accuracy of information provided.

      vi. *Error correction*. It tests if users can recover from mistakes easily and if they make mistakes easily due to system's design.

   (d) Learnability. It evaluates how easily users can learn to use the system.

3. **Saracevic [42]** analyzes 80 evaluation studies taken from the *object* literature[2]. As a result, he proposes a framework to classify the studies. In this framework, usability encompasses the criteria: *Content*, *Process*, *Format* and *Overall assessment*, which are composed of the sub-criteria summarized in Table 2.

4. **Snead et al. [46]** distinguishes between usability and accessibility:

   (a) *Usability*. It determines the extent to which a DL, in whole or in part, enables users to intuitively use its features. It encompasses the sub–criteria:

[2]The difference between *meta* and *object* literature was presented in section 1.

i. *Navigation*: ability to traverse a site using available navigation site tools (e.g., back buttons, links...).
ii. *Content presentation*: the content is presented in a logical manner that is clear and easy to understand.
iii. *Labels*: toolbars, buttons, icons, drop–down features are sensibly presented and labeled.
iv. *Search process*: search features enhance location and retrieval of relevant materials.

(b) *Accessibility*. It determines the extent to which a DL, in whole or in part, provides users with disabilities the ability to interact with the DL. It encompasses the sub–criteria:
i. *Alternate forms of content*: users with visual or auditory disabilities are given access to all content through provision of alternate, equivalent formats.
ii. *Color independent*: users with color deficits and other visual disabilities can access all content (i.e., the DL site does not rely on specific color to convey content).

5. **Tsakonas et al. [49, 50]** propose a usability evaluation similar to Jeng's model. As summarized in Table 2, the main differences are related to the sub-criteria organization.

6. **Xie [52]** conducts an experiment where users are instructed to develop a set of criteria for DL evaluation. The result regarding usability is summarized in Table 2.

3.2 Usefulness

Similarly to usability, most authors consider usefulness as a complex concept composed of several criteria. Table 3 sums up the criteria proposed by Saracevic [42], Tsakonas et al. [49, 50] and Xie [52].

1. As a result of an experimental study, **Xie [52]** identifies the following criteria:

(a) *Scope*. DL scope has to be clearly defined, so users can immediately judge whether they have accessed the right DL.
(b) *Authority*. Authority control is the practice of creating and maintaining index terms for bibliographic material. It enables cataloguers to disambiguate items with similar or identical headings (e.g., two authors who happen to have published under the same name can be distinguished from each other by adding middle initials, a descriptive epithet to the heading of both authors...). In addition, authority control is used to collocate materials that logically belong together, although they present themselves differently (e.g., authority records are used to establish uniform titles, which can collocate all versions of a given work together even when they are issued under different titles).
(c) *Accuracy*. If information is inaccurate, there is no reason for people to use it.
(d) *Completeness*. A good DL covers its subjects thoroughly and is able to provide information that meets the demands of users with varying levels of information need.
(e) *Currency*. DL content should be updated frequently.

2. **Saracevic's [42]** framework does not define explicitly usefulness criteria. Although the criteria summarized in Table 3 are originally included as *usability criteria*, we have decided to reclassify them as *usefulness criteria* to facilitate the comparison of Saracevic's framework with other evaluation proposals.

3. **Tsakonas et al. [49]** differentiates between *goal* and *resource* criteria.

(a) *Goal criteria* are *relevance* (topical relevance, commitment with the quality of information), *utility* and *complexity*.

Author	Criterion	Sub-criterion
Evans et al. [11]	Visibility of system status	
	Match between system and the real world	
	User control and freedom	
	Consistency and standards	
	Error prevention	
	Recognition	
	Flexibility and efficiency of use	
	Aesthetic and minimalist design	
	Help users recignize, diagnose, and recover from errors	
	Help and documentation	
Jeng [23, 24]	Efectiveness	
	Efficiency	
	Satisfaction	Ease of use Organization of information Labeling Visual appearance Content Error correction
	Learnability	
Saracevic [42]	Content	Accessibly, availability Clarity (as presented) Complexity (organization, structure) Understanding, effort to understand
	Process	Learnability to carry out Effort/time to carry out Convenience, ease of use Lostness (confusion) Support for carrying out Completion (achievement of task) Interpretation difficulty Sureness in results Error rate
	Format	Attractiveness Sustaining efforts Convenience, ease of use Consistency Representation of labels Communicativeness of messages
	Overall assessment	Satisfaction Success Relevance, usefulness of results Impact, value Quality of experience Barriers, irritability Preferences Learning
Snead et al. [46]	Usability	Navigation Content presentation Labels Search process
	Accessibility	Alternate forms of content Color independent
Tsakonas et al. [49, 50]	Effectiveness	User performance Error generation
	Efficiency	Completion time Learnability Task completion context
	Satisfaction	Aesthetic comfort Readability
Xie [52]	Interface usability	Search and browse Navigation Help features View and output Accessibility
	User opinion solicitation	User satisfaction User feedback Contact information

Table 2: Usability criteria

(b) *Resource criteria* are *currency*, *level of information* (users' information searching behavior has demonstrated that despite retrieval of full text resources is significant, other levels of information, such as abstracts, are also preferred), *reliability* and *format*.

In addition, in [50] Tsakonas et al. consider *coverage* of the deposited documents as an important usefulness criterion.

Author	Criterion	Sub-criterion
Saracevic [42]	Content	Informativeness Transparency Adequacy Coverage, overlap Quality, accuracy Validity, reliability Authority
Tsakonas et al. [49, 50]	Goal sub-criteria	Relevance Utility Complexity Coverage
	Resource sub-criteria	Currency Level of information Reliability Format
Xie [52]	Scope	
	Authority	
	Accuracy	
	Completeness	
	Currency	

Table 3: Usefulness criteria

3.3 Criteria prioritization

Several works try to identify which evaluation criteria are the most the important from the users' perspective. As we will see, there is a lack of consensus on this issue yet.

An experiment conducted by Kani-Zahibi et al. [27] shows that *finding information easily and quickly in DLs* and *being able to be easily familiarized with DLs* are the two most important DL requirements. In addition, the experiment shows that *supporting collaborative knowledge working* is a minor requirement, contradicting the opinion of Blandford et al. [5].

Xie reports in [53] that *interface usability* and *system performance* are the most important criteria. However, in a previous work [52] she reported that the most relevant criteria was *interface usability* and *collection quality*.

End–user opinion alone is not enough to evaluate a DL. Other stakeholder profiles should be considered. Zhang [54] reports an experiment where different groups of stakeholders (end–users, librarians, DL developers, DL administrators and researchers) are asked to prioritize evaluation criteria for DLs. The research identifies a divergence among the stakeholder groups regarding what criteria should be used for DL evaluation. In the experiment, 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. According to Zhang, the underlying reason for the lowest agreement on technology evaluation is presumably associated with the end–users and librarians unfamiliarity with technological issues. 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.

In [40], Quijano–Solis et al. carry out an experiment to register changes in users' perceptions of the main characteristics and preferred search options in DLs. In the experiment, users are asked to answer a first questionnaire related to criteria importance. Afterwards, they make some tasks to become familiar with a DL. Finally, users answer a second questionnaire analogous with the first one. The result of the experiment shows great changes in the users' opinion. For instance, in the second questionnaire 65% of the participants said that *searching by title* was the preferred way to get information from DLs, contrasting with 39% that marked that option in the first questionnaire. According to Quijano–Solis, more research should be done to understand the nature of those changes, including their randomness.

Garibay et al. [15] propose to use the Kano model [28] to re–prioritize evaluation criteria to take into account the relation between the criteria satisfaction perceived by the users compared to the satisfaction level that they would desire the DL had. In order to adjust the importance of each criterion, equation 1 is used, where:

1. imp_{adj} stands for the adjusted importance of the criterion.
2. imp_0 stands for the importance the criterion has according to the DL users.
3. s_0 stands for how much the DL is currently satisfying the criterion according to the users' opinion.
4. s_1 stands for how much the DL should satisfy the criterion according to the users' opinion.
5. k is the *Kano parameter*. Kano model categorizes the attributes of a product or service based on how well they are able to satisfy customer needs. The model uses three categories, each one with a different k value set by an expert team. The Kano categories are:
   (a) *One–dimensional attributes* (performance needs) are typically what we get by just asking customers what they want. These requirements satisfy (or dissatisfy) in proportion to their presence in the product or service. High performance of a product leads to high customer satisfaction.
   (b) *Attractive attributes* (excitement needs). Absence does not cause dissatisfaction because they are not expected by customers; therefore customers are unaware of what they are missing. However, achievement of these attributes delights the customer, and satisfaction increases with increasing attribute performance.
   (c) *Must–be attributes* (basic needs). Customers take them for granted when fulfilled. However, if the product or service does not meet the basic needs sufficiently, the customer will become very dissatisfied.

$$\text{imp}_{\text{adj}} = \text{imp}_0 \times \left(\frac{s_1}{s_0}\right)^{\frac{1}{k}} \tag{1}$$

As a matter of example, suppose users value from 1 to 5 the criteria importance and how much the DL satisfies them. Initially, users think the importance of a certain criterion c is 4 (i.e., $\text{imp}_0 = 4$) and the DL satisfies the criterion at a level of 3 (i.e., $s_0 = 3$). However, users think the level of satisfaction should be 4 (i.e., $s_1 = 4$). Imagine c falls into the category of *attractive attributes*, which has been valuated with $k = 2$ by the expert team. Hence, the adjusted importance is 4.618 (see Equation 2).

$$\text{imp}_{\text{adj}} = 4 \times \left(\frac{4}{3}\right)^{\frac{1}{2}} = 4.618 \tag{2}$$

3.4 Criteria correlation

In order to minimize the number of criteria that have to be taken into account to evaluate a DL, several authors have analyzed the possible correlation among the criteria.

According to Tsakonas et al. [49], there is a correlation between usefulness and usability. In addition, Jeng [24], Marchionini [34] and Tsakonas et al. [49] have identified the intra-usability relations summarized in Table 4.

1. Using the ANalysis Of VAriance (ANOVA), Jeng [24] concludes that there exist interlocking relationships among *effectiveness*, *efficiency*, and *satisfaction*.

2. Marchionini [34] reports that there is a positive correlation between *system interface* and *learning impact*. In addition, he notes that there is a lack of correlation between *demographics* and *learning impact*.

3. Tsakonas reports correlations between (i) *ease of use* and *navigation*, (ii) *ease of use* and *learnability*, (iii) *navigation* and *aesthetics*, and (iv) *terminology* and *learnability*.

Author	Criterion
Jeng [24]	Effectiveness ↔ Efficiency
	Efficiency ↔ Satisfaction
	Satisfaction ↔ Effectiveness
Marchionini [34]	Learnability ↔ System interface
Tsakonas et al. [50]	Ease of use ↔ Navigation
	Ease of use ↔ Learnability
	Navigation ↔ Aesthetics
	Terminology ↔ Learnability

Table 4: Correlations between usability criteria

Table 5 summarizes the correlations found by Tsakonas et al. [49] among usefulness criteria: between (i) *reliability* and *format*, (ii) *reliability* and *level of information*, and (iii) *coverage* and *level of information*.

Author	Criterion
Tsakonas et al. [50]	Reliability ↔ Format
	Reliability ↔ Level of information
	Level of information ↔ Coverage

Table 5: Correlations between usefulness criteria

4 Criteria measurement and analysis

Criteria refer to chosen standards to judge things by. Criteria are then used to develop *measures* [42]. For instance, *relevance* is a criterion, *precision* and *recall* are measures, and *human relevance judgment* is a measuring instrument.

As noted by Marchionini [34], the literature sometimes bristles with debates over basic approaches to evaluation, especially with respect to *qualitative* versus *quantitative* measures. According to Blandford et al. [3, 4], quantitative approaches (typically involving controlled studies) can be useful to understand the effects of small but meaningful changes on the design of DLs. On the other hand, qualitative methods, whether applied within a laboratory setting (e.g. think-aloud protocols) or in the users' context of work, can be used in a more exploratory way to identify factors for success. Nowadays, there is an increasing

trend of blending quantitative and qualitative data within a study to provide a broader, deeper perspective. This approach is called *triangulation* [32, 34].

Regarding the measuring instruments to get the data for DL evaluation, two main approaches are followed:

1. *Automated techniques*. In this category we can include the analysis of transaction logs [26, 29, 43], a widely used technique that examines the activity of users in a given time session. A more complex approach is proposed by Moreira et al. [37], who have developed the tool *5SQual*. Such tool is grounded in the formal model *5S* for DLs [16, 17] and it does not only detect problems, but it also suggests possible improvements.

2. *Techniques that require users' participation*. Automated techniques have been criticized for their lack of ability to produce interpretable and qualitative data that help evaluators understand the nature of the usability problem and the impact it has on the user interaction [21]. So, interviews and, especially, questionnaires are the prime methods for collecting qualitative data [1, 3, 12, 13, 15, 20, 23, 24, 25, 27, 32, 40, 48, 50, 52, 53, 54, 55]. In addition, observations where users' actions are recorded are also common methods to get data [9, 30, 34, 47].

In questionnaires, users often express their opinion by means of linguistic assessments instead of numerical values. Although the linguistic approach is less precise than the numerical one, it has the following advantages [7]:

1. The linguistic description is easily understood by human beings even when the concepts are abstract or the context is changing.

2. It diminishes the effects of noise since the more refined an assessment scale is, the more sensitive to noise is (linguistic scales are less refined than numerical scales and consequently they are less sensitive to error apparition and propagation).

The following subsections sum up two alternative approaches used in the context of DL evaluation to compute linguistic measurements: the Likert scales and the aggregation of linguistic information based on a symbolic approach.

4.1 Likert scales

Likert scales [33] provide a range of responses to a given question or statement. In the DL evaluation literature, Likert scales usually include 5 categories of response [15, 24, 50, 53]. However, some authors advocate the usage of 6 categories [54] and 7 categories [12] to add additional granularity.

For instance, suppose the evaluation of a DL regarding the usability criterion *user opinion solicitation* proposed by Xie [52]. Table 6 summarizes the users' assessment with a scale of 5 categories: {VL=Very Low, L=Low, M=Medium, H=High, VH=Very High}.

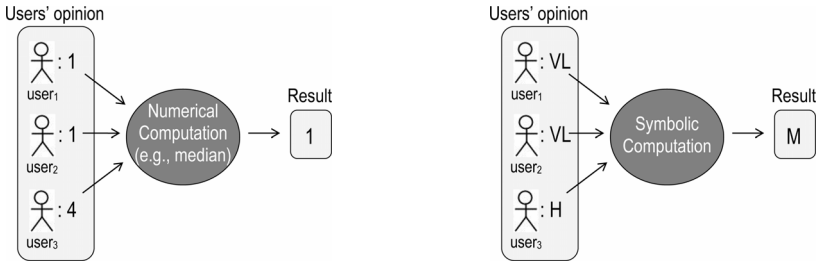
Criterion	$user_1$	$user_2$	$user_3$
User satisfaction	H	VH	VH
User feedback	L	VL	L
Contact information	VL	VL	H

Table 6: Users' responses for the *user opinion solicitation* subcriteria

In order to conclude an evaluation, users' responses are computed numerically (see Figure 3.a). To do so, each level on the scale is assigned to a numeric value, usually starting at 1 and incremented by one for each level. So,

$$\{\mathsf{VL} = 1, \mathsf{L} = 2, \mathsf{M} = 3, \mathsf{H} = 4, \mathsf{VH} = 5\}$$

Many of the authors who use Likert scales for DL evaluation treat individual responses as interval data [12, 15, 24, 53, 54]. Thus, they use the *mean* as the measure of central tendency[3] and the *standard deviation* to measure how much each data value deviates from the mean. For instance, Table 7 summarizes the mean and standard deviation of the data presented in Table 6.



(a) Likert scales (b) LOWA/LWA operators

Figure 3: Alternative approaches to compute users' opinion

Nevertheless, as Blaikie [2] points out, Likert scales fall within the ordinal level of measurement. That is, the response categories have a rank order, but the intervals between values cannot be presumed equal. For instance, the intensity of feeling between VL and L may not be equivalent to the intensity of feeling between other consecutive categories on the Likert scale. The legitimacy of assuming an interval scale for Likert type categories is an important issue, because the appropriate descriptive and inferential statistics differ for ordinal and interval variables. If the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion about the significance of his research [22]. Unfortunately, in [12, 15, 24, 53, 54] no statement is made about the assumption of interval status for Likert data, and no argument made to support it. According to standard statistical texts [2, 10], for ordinal data (i) the *median* and the *mode* are the typical measures of central tendency and (ii) the *range* and the *interquartile range* to measure the data dispersion (see Table 7).

Criterion	**Likert (interval data)**		**Likert (ordinal data)**				**LOWA**
	Mean	*Standard deviation*	*Median*	*Mode*	*Range*	*Interquartile range*	
User satisfaction	4.66	0.44	5	5	1	0.5	VH
User feedback	1.66	0.44	2	2	1	0.5	L
Contact information	2	1.33	1	1	3	1.5	M

Table 7: Analysis of users' responses for the *user opinion solicitation* subcriteria

Table 7 illustrates two important features of the standard measures of central tendency that should be taken into account when analyzing users' opinion:

1. The mean is usually out of the original label set (e.g., *user satisfaction* has a mean of 4.66). In those cases, it is necessary to map the mean values to their corresponding labels (e.g., the *round* function may be used, converting 4.66 into 5, i.e., VH). The median may have the same problem because when there is an even number of users, the median will be calculated as the mean of two middle users' opinions.

2. Whereas the mean is influenced by outliers, the median and the mode are not. As a result, the median and the mode are not always intermediate values. In the example, two users think the satisfaction of the criterion *contact information* is VL and another one thinks it is H. Clearly, the intermediate value falls between VL and H. However, the median and the mode are 1 (i.e., VL).

[3]The purpose of central tendency is to determine the single value that identifies the center of a distribution and best represents the entire set of scores.

4.2 Aggregation of linguistic information based on a symbolic approach

In order to evaluate the quality of DLs, Cabrerizo et al. [7] propose the usage of the two following aggregation operators of linguistic information:

1. The Linguistic Ordered Weighted Averaging (LOWA) operator [19] which is used to combine non–weighted linguistic information.

2. The Linguistic Weighted Averaging (LWA) operator [18] which is used to combine weighted linguistic information and is proposed as a generalization of the LOWA operator applied to combine linguistic information provided by information sources with different importance.

In contrast with the numerical computation of Likert scales presented in Section 4.1, the LOWA and LWA operators compute linguistic labels symbolically (see Figure 3.b).

The behavior of the LOWA and LWA operators is parameterized by selecting different fuzzy linguistic quantifiers. Table 7 summarizes the results of applying the LOWA operator to our example with the "most" quantifier (i.e., it reflects what most of the users think about each criterion). Compared to the analysis of Likert scales, the LOWA operator provides the following benefits:

1. It avoids the simplifying assumption of considering that there is the same distance between all labels.

2. It always produces results contained into the set of linguistic labels.

3. It always generates an intermediate value. In the example, the LOWA result for *contact information* is LOWA(VL,VL,H)=M.

Thanks to the LWA operator, it is possible to calculate a total value that blends the assessment of all users and all subcriteria taking into account the importance of each criterion. Table 8 summarizes the importance of the criteria in our example expressed with the set S of linguistic labels (e.g., the importance of *user satisfaction* is *Very High*). Using equation 3, it can be concluded that, according to most of the users, the DL satisfies the criterion *user opinion solicitation* at a high level.

Criterion	Importance
User satisfaction	VH
User feedback	M
Contact information	L

Table 8: Importance of the *user opinion solicitation* subcriteria

$$\text{LWA}(\text{importance}, \text{users opinion}) = \text{LWA}((\text{VH}, \text{VH}), (\text{M}, \text{L}), (\text{L}, \text{M})) = \text{H} \quad (3)$$

5 Discussions and challenges

The importance of evaluating DLs from the user's perspective is well recognized by the DL community. However, research on user–centered evaluation of DLs seems to be in a preliminary stage.

As outlined in Tables 2 and 3, there are plenty of definitions for usability and usefulness. Because of such lack of common lexicon, it is hard to contrast the experimental results obtained by different authors. For instance, section 3 presents the following contradictory results:

- According to an experiment conducted by Kani-Zahibi et al. [27], the criterion *supporting collaborative knowledge working* has low importance. Nevertheless, the results presented by Blandford et al. [5] support a contrary conclusion.

- Jeng's experiment [24] concludes that *learnability* does not correlate with other usability criteria. However, (i) Marchionini [34] reports a positive correlation between *learning impact* and *system interface*, and (ii) Tsakonas et al. [49] report the correlation of *learnability* with *ease of use* and *terminology*.

Those contradictory results may be a consequence of terminological differences among the criteria definitions managed by the authors. In addition, since most of the experiments have been made with a reduced number of subjects (e.g., Jeng [24] uses 41 subjects, Xie [53] uses 19...), the results may be statistically non–significant.

Regarding the measurement and analysis of users' opinion, two different approaches has been summarized in section 4: Likert scales and LOWA/LWA operators. However, both approaches have not been compared yet.

Although Likert scales fall within the ordinal level of measurement, many authors treat individual responses as interval data [12, 15, 24, 53, 54] without no justification about the assumption of interval status for Likert data. As we have noted, if the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion about the significance of his research.

To sum up, the user–centered evaluation of DLs requires to tackle the following challenges:

1. A consensus on standard definitions for usability and usefulness have to be reached.
2. The minimal threshold of subjects to obtain statistically significant results on the importance and correlation of usability and usefulness criteria should be identified.
3. The assumption of interval status for Likert data in the DL context has to be justified.
4. The advantages and drawbacks of Likert scales compared to LOWA/LWA operators have to be identified. We propose Table 9 as a starting point that should be developed in future work. According to this table, LOWA/LWA operators seem to produce better results (the details are presented in sections 4.1 and 4.2). On the other hand, Likert scales support the measurement of the users's opinion dispersion.

Supported features	**Likert**		**LOWA/LWA**
	Interval data	*Ordinal data*	
Results correspond to labels in the original term set	×	✓	✓
Always combine the input values to produce an intermediate output	✓	×	✓
Combination of weighted criteria	×	×	✓
Measure of dispersion	✓	✓	×

Table 9: Supported features of Likert scales and LOWA/LWA operators to analyze users' opinion

6 Conclusions

In this paper, we have revised the state of the art on the user–centered evaluation of DLs by running a structured literature review covering 35 primary studies and outlining the main advances made up to now. As a result, we have summarized what criteria are being used to evaluate DLs, the importance of each criterion and the inter–criteria correlation. We have also provided information about the measures that are derived from those criteria, the measuring instruments to elicit users' opinion and how the measurements are combined to produce a DL evaluation. To finalize, we have identified a number of challenges for

future research mainly related to (i) the standard definition of usability and usefulness criteria, (ii) the minimum necessary requirements to guarantee the validity of experiments on criteria correlation/priorization and (iii) the comparative analysis of existing proposals to get evaluations by combining the information collected via measuring instruments.

References

[1] Cemal Atakan, Dogan Atilgan, Ozlem Bayram, and Sacit Arslantekin. An evaluation of the second survey on electronic databases usage at ankara university digital library. *The Electronic Library*, 26(2):249–259, 2008.

[2] Norman Blaikie. *Analyzing Quantitative Data: From Description to Explanation*. Sage Publications, 2003.

[3] A. Blandford, A. Adams, S. Attfield, G. Buchanan, J. Gow, S. Makri, J. Rimmer, and C. Warwick. The pret a rapporter framework: Evaluating digital libraries from the perspective of information work. *Information Processing and Management*, 44(1):4–21, January 2008.

[4] Ann Blandford. Understanding users' experiences: evaluation of digital libraries. In *WP7 Workshop on the Evaluation of Digital Libraries*, pages 31–34. DELOS Network of Excellence on Digital Libraries, October 2004.

[5] Ann Blandford, Suzette Keith, Iain Connell, and Helen Edwards. Analytical usability evaluation for digital libraries: a case study. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '04, pages 27–36, New York, NY, USA, 2004. ACM.

[6] Christine L. Borgman. *Digital Library Use. Social Practice in Design and Evaluation*, chapter Designing Digital Libraries for Usability. MIT Press, 2003.

[7] F.J. Cabrerizo, J. Lopez-Gijon, A.A. Ruiz, and E. Herrera-Viedma. A model based on fuzzy linguistic information to evaluate the quality of digital libraries. *International Journal of Information Technology and Decision Making*, 9(3):455–472, 2010.

[8] L Candela, D. Castell, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, and H. Schuldt. The delos digital library reference model - foundations for digital libraries. version 0.96, November 2007.

[9] Jason A. Clark. A usability study of the belgian–american research collection: measuring the functionality of a digital library. *OCLC Systems and Services*, 20(3):115–127, 2004.

[10] Frances Clegg. *Simple Statistics: A Course Book for the Social Sciences*. Cambridge University Press, 1983.

[11] Joanne Evans, Andrew O'Dwyer, and Stephan Schneider. Usability evaluation in the context of digital video archives. In *Fourth DELOS Workshop. Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics.*, pages 79–86, 2002.

[12] Enrique Frias-Martinez, Sherry Y. Chen, and Xiaohui Liu. Evaluation of a personalized digital library based on cognitive styles: Adaptivity vs. adaptability. *International Journal of Information Management*, 29(1):48–56, February 2009.

[13] Norbert Fuhr, Preben Hansen, Michael Mabe, Andras Micsik, and Ingeborg Sølvberg. Digital libraries: A generic classification and evaluation scheme. In *5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 187–199. Springer, September 2001.

[14] Norbert Fuhr, Giannis Tsakonas, Trond Aalberg, Maristella Agosti, Preben Hansen, Sarantos Kapidakis, Claus-Peter Klas, Laszlo; Kovas, Monica Landoni, Andras Micsik, Christos Papatheodorou, Carol Peters, and Ingeborg Sølvberg. Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1):21–38, October 2007.

[15] Cecilia Garibay, Humberto Gutierrez, and Arturo Figueroa. Evaluation of a digital library by means of quality function deployment (qfd) and the kano model. *The Journal of Academic Librarianship*, 36(2):125–132, March 2010.

[16] Marcos Andre Goncalves and Edward A. Fox. 5sl: a language for declarative specification and generation of digital libraries. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 263–272, New York, NY, USA, 2002. ACM.

[17] Marcos Andre Goncalves, Edward A. Fox, Layne T. Watson, and Neill A Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transactions on Information Systems*, 22(2):270–312, 2004.

[18] F. Herrera and E. Herrera-Viedma. Aggregation operators for linguistic weighted information. *IEEE Transactions on Systems, Man and Cybernetics*, 27:646–656, 1997.

[19] F. Herrera, E. Herrera-Viedma, and J. L. Verdegay. Direct approach processes in group decision making using linguistic owa operators. *Fuzzy Sets Syst.*, 79:175–190, April 1996.

[20] Alireza Isfandyari-Moghaddam and Behrooz Bayat. Digital libraries in the mirror of the literature: issues and considerations. *The Electronic Library*, 26(6):844–862, 2008.

[21] Melody Y. Ivory and Marti A Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33:470–516, December 2001.

[22] Susan Jamieson. Likert scales: how to (ab)use them. *Medical Education*, 38(12):1217–1218, 2004.

[23] Judy Jeng. Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *International Journal Of Libraries and Information Services*, 55:96–121, 2005.

[24] Judy Jeng. What is usability in the context of the digital library and how can it be measured? *Information Technology and Libraries*, 24(2):47–56, 2005.

[25] Judy Jeng. Evaluation of the new jersey digital highway. *Information Technology and Libraries*, 27(4):17–24, 2008.

[26] Steve Jones, Sally Jo Cunningham, Rodger J. McNab, and Stefan J. Boddie. A transaction log analysis of a digital library. *Int. J. on Digital Libraries*, 3(2):152–169, 2000.

[27] Kani-Zabihi, Elahe, Ghinea, Gheorghita, Chen, and Sherry, Y. Digital libraries: what do users want? *Online Information Review*, 30(4):395–412, 2006.

[28] N. Kano, K. Seraku, F. Takahashi, and S. Tsuji. Attractive quality and must-be quality hinshitsu quality. *Journal of the Japanese Society for Quality Control*, 14(2):39–48, 1984.

[29] Hao-Ren Ke, Rolf Kwakkelaarb, Yu-Min Taic, and Li-Chun Chenc. Exploring behavior of e-journal users in science and technology: transaction log analysis of elsevier's sciencedirect onsite in taiwan. *Library and Information Science Research*, 24:265–291, 2002.

[30] Rekha Kengeri, Cheryl D. Seals, Hope D. Harley, Himabindu P. Reddy, and Edward A. Fox. Usability study of digital libraries: Acm, ieee–cs, ncstrl, ndltd. *Int. J. on Digital Libraries*, 2(2-3):157–169, 1999.

[31] Barbara Kitchenham. Procedures for performing systematic reviews. Technical Report TR/SE-0401, Software Engineering Group. Department of Computer Science. Keele University, 2004.

[32] Patty Kostkova and Gemma Madle. User-centered evaluation model for medical digital libraries. *Lecture Notes In Artificial Intelligence*, pages 92–103, 2009.

[33] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

[34] Gary Marchionini. Evaluating digital libraries: a longitudinal & multifaceted view. library trends, 49 (2): 304-333. *Library Trends*, 49(2):304–333., 2001.

[35] Gary Marchionini, Catherine Plaisant, and Anita Komlodi. The people in digital libraries: Multifaceted approaches to assessing needs and impact. In *In Digital library use: Social practice in design and evaluation*, pages 119–160. MIT Press, 2003.

[36] R. B. Miller. Human ease of use criteria and their tradeoffs. Technical Report IBM Report TR 00.2185, IBM Corporation, 1971.

[37] Barbara L. Moreira, Marcos A. Goncalves, Alberto H. Laender, and Edward A. Fox. Automatic evaluation of digital libraries with 5SQual. *Journal of Informetrics*, 3(2):102–123, April 2009.

[38] Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993.

[39] International Standards Organization. *Ergonomic requirements for office work with visual display terminals*, chapter Part 11: Guidance on usability (ISO DIS 9241-11). International Standards Organization, 1994.

[40] Alvaro Quijano-Solis and Raul Novelo-Pena. Evaluating a monolingual multinational digital library by using usability: An exploratory approach from a developing country. *International Information and Library Review*, 37(4):329–336, 2005.

[41] Tefko Saracevic. Digital library evaluation: toward an evolution of concepts. *Library Trends*, 49(2):350–369, 2000.

[42] Tefko Saracevic. Evaluation of digital libraries: An overview. In *WP7 Workshop on the Evaluation of Digital Libraries*, pages 13–30. DELOS Network of Excellence on Digital Libraries, October 2004.

[43] Michalis Sfakakis and Sarantos Kapidakis. User behavior tendencies on data collections in a digital library. In Maristella Agosti and Costantino Thanos, editors, *ECDL*, volume 2458 of *Lecture Notes in Computer Science*, pages 550–559. Springer, 2002.

[44] Brian Shackel. *Human Factors for Informatics Usability*, chapter Usability – Context, framework, definition, design and evaluation, pages 21–37. Cambridge University Press, 1991.

[45] Alan F. Smeaton and Jamie Callan. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries*, 57(4):299–308, 2005.

[46] John T. Snead, John Carlo Bertot, Paul T. Jaeger, and Charles R. McClure. Developing multi-method, iterative, and user-centered evaluation strategies for digital libraries: Functionality, usability, and accessibility. In *Proceedings of the American Society for Information Science and Technology*, 2005.

[47] A. G. Sutcliffe, M. Ennis, and J. Hu. Evaluating the effectiveness of visual user interfaces for information retrieval. *Int. J. Hum.-Comput. Stud.*, 53:741–763, November 2000.

[48] James Y. L. Thong, Weiyin Hong, and Kar Yan Tam. Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences? *International Journal of Human Computer Studies*, 57(3):215–242, 2002.

[49] Giannis Tsakonas, Sarantos Kapidakis, and Christos Papatheodorou. Evaluation of user interaction in digital libraries. In *WP7 Workshop on the Evaluation of Digital Libraries*, pages 45–60. DELOS Network of Excellence on Digital Libraries, October 2004.

[50] Giannis Tsakonas and Christos Papatheodorou. Exploring usefulness and usability in the evaluation of open access digital libraries. *Information Processing & Management*, 44(3):1234–1250, May 2008.

[51] J Webster and R T Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):13–23, 2002.

[52] Hong Iris Xie. Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library and information Science Research*, 28:433–452, 2006.

[53] Hong Iris Xie. Users' evaluation of digital libraries (dls): Their uses, their criteria, and their assessment. *Information Processing and Management*, 44(3):1346–1373, May 2008.

[54] Ying Zhang. Developing a holistic model for digital library evaluation. *J. Am. Soc. Inf. Sci. Technol.*, 61:88–110, January 2010.

[55] Alesia Zuccala and Charles Oppenheim. Managing and evaluating digital repositories. *Information Research*, 13(1), March 2008.