A review of quality evaluation of digital libraries based on users' perceptions
Rubén Heradio, David Fernández-Amorós, Francisco Javier Cabrerizo and Enrique Herrera-Viedma

Journal of Information Science 2012 38: 269 originally published online 19 April 2012
DOI: 10.1177/0165551512438359

The online version of this article can be found at:
http://jis.sagepub.com/content/38/3/269

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
CILIP
Chartered Institute of Library and Information Professionals

Additional services and information for Journal of Information Science can be found at:

Email Alerts: http://jis.sagepub.com/cgi/alerts

Subscriptions: http://jis.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jun 12, 2012

OnlineFirst Version of Record - Apr 19, 2012

What is This?
A review of quality evaluation of digital libraries based on users’ perceptions

Rubén Heradio
Distance Learning University of Spain (UNED), Spain

David Fernández-Amorós
Distance Learning University of Spain (UNED), Spain

Francisco Javier Cabrerizo
Distance Learning University of Spain (UNED), Spain

Enrique Herrera-Viedma
University of Granada, Spain

Abstract
In the past two decades, the use of digital libraries (DLs) has grown significantly. Accordingly, questions about the utility, usability and cost of DLs have started to arise, and greater attention is being paid to the quality 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 the user’s opinion. The literature on this topic has produced a set of varied criteria to judge DLs from the user’s perspective, measuring instruments to elicit users’ opinions, and approaches to analyse the elicited data to conclude an evaluation. This paper provides a literature review of the quality evaluation of DLs based on users’ perceptions. Its main contribution is to bring together previously disparate streams of work to help shed light on this thriving area. In addition, the various studies are discussed, and some challenges to be faced in the future are proposed.

Keywords
digital libraries (DLs); quality evaluation; user’s satisfaction

1. Introduction
A digital library (DL) is a collection of information that has associated services delivered to user communities using a variety of technologies [1]. In general, DLs are the logical extension of physical libraries [2–5] in an electronic information society. Such extensions offer new levels of access to broader audiences of users [6, 7].

The use of DLs has grown significantly in the past two decades [8]. At the end of the 1980s, DLs were barely a part of the landscape of librarianship, information science or computer science. A decade later, by the end of the 1990s, research, practical developments and general interest in DLs had exploded globally. The accelerated growth of numerous and highly varied efforts related to DLs has continued unabated in the 2000s [9, 10]. Since the 1990s, the internet and the web have become the primary platform for libraries to build and deliver information resources, services and instructions. Library users are now offered a variety of resources with different forms of interactivity, and with different levels of media richness. They can obtain research data and publications as needed, without the massive investment of capital and infrastructure to house vast physical collections. Information seeking in DLs has become an indispensable tool in...
academia, and personal use is increasing every day [11]. Therefore there are a large number of users of DLs whose
expectations and demands for better service and functionality are increasing.

Once the importance and applicability of this type of information system have been definitively established, questions
about the utility, usability and cost of DLs have started to arise. Defining 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 [12].
This has led to the expansion of DL evaluation to sectors such as database structure, network architecture, protocol inter-
operability, the development of intelligent and adaptive technologies, the performance of retrieval algorithms, collection
development, digitization policy assessment, usability, information architecture, interaction design, information beha-
vour and many others [13, 14].

As the final aim of a DL system is to enable people to access human knowledge at any time and anywhere, in a
friendly multimodal way, by overcoming barriers of distance, language and culture, and by using multiple network-
connected devices [15], the quality of DLs needs to be judged by their users. DLs are intended to serve users; if these
systems are not used, they fall into oblivion and terminate their operation [16]. Therefore one of the main aspects to be
considered in DL evaluation is the user’s perspective, determining the extent to which the DL addresses the real needs
of its users [17].

User-centred evaluation of DLs has drawn considerable attention during recent years [18]. Research in this area has
produced a set of varied criteria by which to judge DLs from the user’s perspective, measuring instruments to elicit users’
opinions, and approaches to analyse the elicited data to conclude an evaluation. However, despite the importance of
the quality evaluation of DLs, few literature reviews have been performed in this field according to users’ judgments, which
is necessary to bring together previously disparate streams of work and help shed light on this thriving area.

The aim of this paper is to provide a literature review of the quality evaluation of DLs, based on the user’s perspective.
According to Saracevic [19], the literature on DL evaluation can be divided into two distinct types: (1) meta or ‘about’
literature (i.e., works that suggest evaluation concepts, models, approaches and 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-litera-
ture on user-centred evaluation of DLs. Its purpose is to create a firm foundation for advancing knowledge that will facil-
itate theory development, close areas where a plethora of research exists, and uncover areas where research is needed. In
order to fulfill such goals, our review follows the rigorous and auditable methodology proposed by Kitchenham [20] and
Webster et al. [21]. Specifically, this paper addresses the following research questions:

1. What criteria are proposed to evaluate DLs in a user-centred fashion?
2. How are those criteria measured and processed?
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 the criteria that are proposed for the user-centred 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’ opinions, 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 followed a systematic and structured method inspired by the guidelines of Kitchenham [20]
and Webster et al. [21]. Below, we detail the main data regarding the review process and its structure.

The aim of this paper is to review the quality evaluation of DLs based on users’ perceptions. The term user has varied
meanings in the DL context. For instance, the DELOS Digital Library Reference Model [22] identifies the following
types of actor 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:
   Content consumers are the purchasers of the DL content.
   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.
   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.

Journal of Information Science, 38 (3) 2012, pp. 269–283 © The Author(s), DOI: 10.1177/0165551512438359
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 that the available resources impose on the definition of the DL.

4. **DL applications 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 quality evaluation of DLs from the content consumer’s point of view. Taking it into account, the aim of this review is to answer the following research questions (RQs):

- **RQ1**: What criteria are used to evaluate the quality of DLs in a user-centred fashion?

This question motivates the following subquestions:

- Do all criteria have the same importance in the evaluation?
- Is there any correlation among the criteria?
- **RQ2**: How are those criteria measured?

This question motivates the following subquestions:

- Are the measures quantitative or qualitative?
- What instruments are used to elicit users’ opinions?
- How are the measurements analysed?
- **RQ3**: What are the challenges to be faced in the future?

Sections 3–5 attempt to answer questions RQ1, RQ2 and RQ3 respectively. To do so, as recommended by Webster et al. [21], we have used both manual and automated methods to make a selection of candidate papers in leading journals, conferences and other related events. Table 1 classifies the primary studies [20] we have reviewed according to the year and type of publication. Of the 41 papers included in the review, 26 were published in journals, 8 in conferences, 4 in workshops, 2 in books, and 1 as a technical report.

### 3. Criteria for quality evaluation of DLs

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Journals</th>
<th>Conferences</th>
<th>Workshops</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>[23]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>[24]</td>
<td>[25], [26]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>[8]</td>
<td></td>
<td>[27]</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td>[16]</td>
</tr>
<tr>
<td>2004</td>
<td>[28]</td>
<td>[29]</td>
<td>[13], [19], [30]</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>[31], [32], [33]</td>
<td>[34]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>[35], [36]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>[15], [37]</td>
<td></td>
<td></td>
<td>[22]</td>
</tr>
<tr>
<td>2008</td>
<td>[17], [38], [39], [40], [41], [42], [43], [44]</td>
<td>[47]</td>
<td></td>
<td>[48]</td>
</tr>
<tr>
<td>2009</td>
<td>[45], [46]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>[10], [18], [49], [50]</td>
<td>[51], [52]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>[14]</td>
<td></td>
<td></td>
<td>[53]</td>
</tr>
</tbody>
</table>

*Journal of Information Science, 38 (3) 2012, pp. 269–283 © The Author(s), DOI: 10.1177/0165551512438359*
Fuhr et al. [25] 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.

Information retrieval provides the standard measures of precision and recall to evaluate the quality of the system–content interaction; however, their applicability to evaluation of the DL user experience has been called into question. So, to Fuhr et al. [15], these measures have not been defined to ‘model user satisfaction, result pertinence or system effectiveness for a given task content’, and instead they advocate the development of DL-specific evaluation schemes. According to [25], it is not the precision and recall measures that are inappropriate, but rather the benchmarking collections used in competitions such as TREC that ‘lack the rich structure and inter-document relationships that are typical for DL collections’.

Although Fuhr et al.’s model addresses system-centred evaluations [15], it has remarkably influenced several user-centred proposals. In particular, Tsakonas et al. [13] propose a user-centred model focused on the relations between the components of Fuhr’s model. These relations are shown in Figure 1.

- The content–system pair is related to performance criteria (precision, recall, response time, etc.).
- The user–system pair is related to the usability 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 that supports exploitation of all the available functionalities. A usable system is easy to learn, is 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’s tasks and needs.

This paper is focused on the user–system and the user–content interactions (see the shadowed area in Figure 1). At present, 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.

### 3.1. Usability and usefulness

According to Shackel [54], the definition of informatics usability was probably first attempted in 1971 by Miller [55] in terms of measures for ‘ease of use’. Since then, a wide variety of definitions for informatics usability have been proposed: for example, Jeng [32] reviews 15 different definitions. In this paper, we restrict our usability review to the DL context.

Most authors consider usability as a complex concept, composed of several criteria. Figure 2 outlines the usability criteria and subcriteria considered by Evans et al. [27], Jeng [31, 32, 48], Saracevic [19], Snead et al. [34], Tsakonas et al. [13, 42], and Xie [35]:

![Figure 1](image-url) **Figure 1.** Digital library evaluation model proposed by Tsakonas et al. [13].
Evans et al. [27] propose a usability evaluation framework that takes into account the following criteria:

1. **Visibility of system status.** The system should always keep users informed about what is going on, through appropriate feedback within a reasonable time.
2. **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.
3. **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.
4. **Consistency and standards.** Users should not have to wonder whether different words, situations, or actions mean the same thing.
5. **Error prevention.** Even better than good error messages is a careful design that prevents a problem from occurring in the first place.

6. **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.

7. **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.

8. **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.

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

10. **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, be focused on the user’s task, list concrete steps to be carried out, and not be too large.

Jeng [31, 32, 48] proposes an evaluation model that applies the usability definition of ISO 9241-11 [56]. It examines the following criteria:

1. **Effectiveness.** It evaluates whether the system can provide information and functionality effectively.

2. **Efficiency.** It evaluates whether the system can be used to retrieve information efficiently.

3. **Satisfaction.** This encompasses the following subcriteria:
   - **Ease of use.** It evaluates the user’s perception of the ease of use of the system.
   - **Organization of information.** It evaluates whether the system’s structure, layout, and organization meet the user’s satisfaction.
   - **Labelling.** It evaluates from the user’s perception whether the system provides clear labelling, and whether the terminology used is easy to understand.
   - **Visual appearance.** It evaluates the site’s design to see whether it is visually attractive.
   - **Contents.** It evaluates the authority and accuracy of the information provided.
   - **Error correction.** It tests whether users can recover from mistakes easily, and whether they make mistakes easily because of the system’s design.

4. **Learnability.** It evaluates how easily users can learn to use the system.

Saracevic [19] analyses 80 evaluation studies taken from the object literature. 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 subcriteria summarized in Figure 2.

Snead et al. [34] distinguish between usability and accessibility:

1. **Usability.** This determines the extent to which a DL, in whole or in part, enables users to use its features intuitively. It encompasses the following subcriteria:
   - **Navigation:** the ability to traverse a site using available navigation site tools (e.g., back buttons, links, etc.).
   - **Content presentation:** the content is presented in a logical manner that is clear and easy to understand.
   - **Labels:** toolbars, buttons, icons, drop-down features are sensibly presented and labelled.
   - **Search process:** search features enhance location and retrieval of relevant materials.

2. **Accessibility.** This determines the extent to which a DL, in whole or in part, enables users with disabilities to interact with the DL. It encompasses the following subcriteria:
   - **Alternative forms of content:** users with visual or auditory disabilities are given access to all content through the provision of alternative, equivalent formats.
   - **Colour independent:** users with colour deficits and other visual disabilities can access all content (i.e., the DL site does not rely on specific colour to convey content).

Tsakonas et al. [13, 42] propose a usability evaluation similar to Jeng’s model. As summarized in Figure 2, the main differences are related to the subcriteria organization.

Xie [35] conducts an experiment in which users are instructed to develop a set of criteria for DL evaluation. The result regarding usability is summarized in Figure 2.
Similarly to usability, most authors consider usefulness as a complex concept composed of several criteria. Figure 2 also sums up the criteria proposed by Xie [35], Saracevic [19] and Tsakonas et al. [13, 42]. As a result of an experimental study, Xie [35] identifies the following criteria:

1. **Scope.** DL scope has to be clearly defined, so that users can immediately judge whether they have accessed the right DL.
2. **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, etc.). 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).
3. **Accuracy.** If information is inaccurate, there is no reason for people to use it.
4. **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.
5. **Currency.** DL content should be updated frequently.

Saracevic’s [19] framework does not define usefulness criteria explicitly. Although the criteria summarized in Figure 2 are originally included as usability criteria, we have decided to reclassify them as usefulness criteria, to facilitate comparison of Saracevic’s framework with other evaluation proposals.

Tsakonas et al. [13] differentiate between goal and resource criteria:

1. **Goal criteria** are relevance (topical relevance, commitment with the quality of information), utility and complexity.
2. **Resource criteria** are currency, level of information (users’ information searching behaviour has demonstrated that, although retrieval of full text resources is significant, other levels of information, such as abstracts, are also preferred), reliability and format.

In addition, in [42], Tsakonas et al. consider coverage of the deposited documents as an important usefulness criterion.

### 3.2. Criteria prioritization

Several works try to identify which evaluation criteria are the most important from the user’s perspective. As we shall see, there is as yet a lack of consensus on this issue.

An experiment conducted by Kani-Zahibi et al. [36] 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. [29]. Xie reports in [43] that interface usability and system performance are the most important criteria. However, in a previous work [35], she reported that the most relevant criteria were interface usability and collection quality.

Zhang [18] reports an experiment in which different groups of stakeholders (end-users, librarians, DL developers, DL administrators and researchers) were 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 the indirect relationship between DL use and context might be the two factors causing the larger divergence for content and context evaluation criteria.

In [33], Quijano-Solis et al. describe an experiment to register changes in users’ perceptions of the main characteristics and preferred search options in DLs. In the experiment, users were first asked to answer a questionnaire related to criteria importance. Then they performed some tasks to become familiar with a DL. Finally, they answered a second questionnaire analogous to the first one. The result of the experiment shows great changes in the users’ opinions.
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% who marked that option in the first questionnaire. According to Quijano-Solis et al. [33], more research should be done to understand the nature of these changes, including their randomness.

Garibay et al. [50] propose using the Kano model [57] to reprioritize evaluation criteria to take into account the relation between the criteria satisfaction perceived by the users compared with the satisfaction level that they would desire the DL to have. In order to adjust the importance of each criterion, the following equation is used:

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

(1)

where \( \text{imp}_{\text{adj}} \) is the adjusted importance of the criterion; \( \text{imp}_0 \) is the importance the criterion has according to the DL users; \( s_0 \) indicates how much the DL is currently satisfying the criterion according to the users’ opinion; \( s_1 \) indicates how much the DL should satisfy the criterion according to the users’ opinion; and \( k \) is the Kano parameter. The Kano model categorizes the attributes of a product or service based on how well it is able to satisfy customer needs.

The model uses three categories, each with a different \( k \) value set by an expert team. The Kano categories are as follows:

- **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.

- **Attractive attributes** (excitement needs). Their 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.

- **Must-be attributes** (basic needs). Customers take these for granted when they are fulfilled. However, if the product or service does not meet the basic needs sufficiently, the customer will become very dissatisfied.

For 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 evaluated with \( k = 2 \) by the expert team. Hence the adjusted importance is 4.618:

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

**3.3. Criteria correlation**

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

According to Tsakonas et al. [13], there is a correlation between usefulness and usability. In addition, Jeng [32], Marchionini [24] and Tsakonas and Papatheodorou [42] have identified the following intra-usability and intra-usefulness relations, which are summarized in Table 2. Using Analysis of Variance (ANOVA), Jeng [32] concludes that there exist interlocking relationships among effectiveness, efficiency and satisfaction. Marchionini [24] 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.

Tsakonas and Papatheodorou [42] report the following correlations:

1. **Usability criteria:** between (i) ease of use and navigation, (ii) ease of use and learnability, (iii) navigation and aesthetics, and (iv) terminology and learnability.

2. **Usefulness criteria:** between (i) reliability and format, (ii) reliability and level of information, and (iii) coverage and level of information.
4. Criteria measurement and analysis

As noted by Marchionini [24], the literature sometimes bristles with debates over basic approaches to evaluation, especially with respect to qualitative versus quantitative measures. According to Blandford et al. [30, 39], quantitative approaches (typically involving controlled studies) can be useful in understanding 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 user’s context of work, can be used in a more exploratory way to identify factors for success. There is an increasing trend to blend quantitative and qualitative data within a study to provide a broader and deeper perspective. This approach is called triangulation [24, 46].

In obtaining the data for quality evaluation of DLs, automated techniques have been used. In this category we can include the analysis of transaction logs [37, 58–60], a widely used technique that examines the activity of users in a given time session, and the 5SQual tool, a more complex approach proposed by Moreira et al. [12]. Such a tool is grounded in the formal model 5S for DLs [61, 62]. However, 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 [26]. Furthermore, as mentioned in Section 1, since DLs are destined to serve user communities, the user’s opinion is one of the main aspects to be considered in the quality evaluation of DLs. For these reasons, techniques that require the user’s participation such as interviews and questionnaires are the focus of this review, and the prime methods for collecting qualitative data [8, 18, 25, 31–33, 35, 36, 38–47, 49–53]. In addition, observations in which user actions are recorded are also common methods for obtaining data [23, 24, 28, 63].

In these methods, to obtain the users’ quality assessment, they are invited to fill in a survey built on the set of criteria. To measure quality, conventional measurement tools used by the users are devised on cardinal or ordinal scales (Likert scales [64]). However, the scores do not necessarily represent the user’s preference. This is because respondents have to internally convert preference to scores, and the conversion may introduce distortion of the preference [65]. For this reason, a recent method proposes using ordinal fuzzy linguistic modelling [66, 67] to represent the user’s perceptions, and computing tools with words based on the linguistic aggregation operators LOWA [66] and LWA [67] to compute the quality assessments.

The following subsections sum up these two alternative approaches, used in the context of DL evaluation to compute quality measurements based on users’ perceptions: Likert scales and fuzzy linguistic modelling.

4.1. Likert scales

A Likert scale [64] is a psychometric scale commonly used in questionnaires, and is the most widely used scale in survey research. When responding to a Likert questionnaire item, respondents are asked to evaluate their level of agreement or disagreement according to any given subjective or objective criteria. To do so, Likert scales provide a range of responses to a given question or statement.

In the DL evaluation literature, Likert scales usually include five levels of response [32, 42, 43, 50]. However, some authors advocate the use of six [18] or even seven levels [45], for additional granularity. For instance, consider the evaluation of a DL regarding the usability criterion user opinion solicitation proposed by Xie [35], which is composed of three subcriteria. Table 3 summarizes the user’s assessment on a scale of five levels: 1 = very low, 2 = low, 3 = medium, 4 = high, and 5 = very high.
In order to conclude an evaluation, the user’s perceptions are computed numerically (see Figure 3a). To do so, each level on the scale is assigned to a numeric value, usually starting at 1 and incremented by 1 for each level.

Many of the authors who use Likert scales for DL evaluation treat individual responses as interval data [18, 32, 43, 45, 50]. Thus they use the mean as the measure of central tendency, and the standard deviation to measure how much each data value deviates from the mean. For instance, Table 4 summarizes the mean and standard deviation of the data presented in Table 3.

Nevertheless, as Blaikie [68] points out, Likert scales fall within the ordinal level of measurement. That is, the response levels have a rank order, but the intervals between values cannot be presumed equal; one cannot assume that respondents perceive all pairs of adjacent levels as equidistant. For instance, the intensity of feeling between ‘very low’ and ‘low’ 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, researchers increase the chance of coming to the wrong conclusion about the significance of their research [69]. Unfortunately [18, 32, 43, 45, 50], no statement is made about the assumption of interval status for Likert data, and no argument is made to support it. According to standard statistical texts [68, 70], for ordinal data (1) the median and the mode are the typical measures of central tendency, and (2) the range and the interquartile range measure the data dispersion (see Table 4).

### 4.2. Fuzzy linguistic modelling

In [47, 49, 51–53], to represent the user’s perceptions of the quality of the DL, fuzzy linguistic information [66, 67, 71–74] is used instead of numerical values. In particular, ordinal fuzzy linguistic modelling [66, 67] is used to represent the user’s perceptions.

#### Table 3. User’s responses for subcriteria of criterion user opinion solicitation using Likert scales

<table>
<thead>
<tr>
<th>Subcriteria</th>
<th>user1</th>
<th>user2</th>
<th>user3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User satisfaction</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>User feedback</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Contact information</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Table 4. Analysis of users’ responses for subcriteria of criterion user opinion solicitation

<table>
<thead>
<tr>
<th>Subcriteria</th>
<th>Likert (interval data)</th>
<th>Likert (ordinal data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>4.66</td>
<td>0.44</td>
</tr>
<tr>
<td>User feedback</td>
<td>1.66</td>
<td>0.44</td>
</tr>
<tr>
<td>Contact information</td>
<td>2</td>
<td>1.33</td>
</tr>
</tbody>
</table>
The ordinal fuzzy linguistic approach is a very useful kind of fuzzy linguistic approach, used for modelling the process of computing with words, as well as the linguistic aspects of problems. It facilitates fuzzy linguistic modelling because it simplifies definition of the semantic and syntactic rules. It is defined by considering a finite and totally ordered label set $S = \{s_i\}, i \in \{0, \ldots, T\}$ in the usual sense, that is, $s_i \geq s_j$ if $i \geq j$, and with odd cardinality. Typical values of cardinality used in linguistic models are odd values, such as 7 or 9, with an upper limit of granularity of 11 or no more than 13, where the mid-term represents an assessment of ‘approximately 0.5’, and the rest of the terms are placed symmetrically around it. The semantics of the linguistic term set is established from the ordered structure of the label set by considering that each linguistic term for the pair $(s_i, s_{T-i})$ is equally informative. For example, we can use the following set of five labels to provide the user’s evaluations: VL = very low, L = low, M = medium, H = high, VH = very high. Advantages of the ordinal fuzzy linguistic approach are the simplicity and rapidity of its computational model. It is based on symbolic computation [66, 67], and acts by direct computation on labels by taking into account the order of such linguistic assessments in the ordered structure of linguistic terms. In order to evaluate the quality of DLs, the two following aggregation operators of linguistic information are used:

1. the linguistic ordered weighted averaging (LOWA) operator [66], which is used to combine non-weighted linguistic information;
2. the linguistic weighted averaging (LWA) operator [67], 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 to the numerical computation of Likert scales presented in Section 4.1, the LOWA and LWA operators compute linguistic labels symbolically, and therefore linguistic approximation processes are unnecessary; this simplifies the processes of computing with words (see Figure 3b).

The behaviour of the LOWA and LWA operators is parameterized by selecting different fuzzy linguistic quantifiers. Table 5 summarizes the results of applying the LOWA operator to our example with the ‘most’ quantifier, defined as $Q(r) = r^{1/2}$ (i.e., it reflects what most of the users think about each criterion). Compared with 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 in the set of linguistic labels.
3. It always generates an intermediate value. In the example, the LOWA result for contact information is $\text{LOWA(VL, VL, H)} = M$.

Furthermore, thanks to the LWA operator, it is possible to calculate a total value that blends the assessment of all users and all criteria, taking into account the importance of each criterion. Table 6 summarizes the importance of the subcriteria in our example, expressed with the set $S$ of linguistic labels (e.g. the importance of user satisfaction is Very High). Since $\text{LWA(Importance, user’s opinion)} = \text{LWA}[\text{VH, VH}, (M, L), (L, M)] = H$, it can be concluded that, according to most of the users, the DL satisfies the criterion user opinion solicitation at a high level.

### Table 5. Users’ responses for subcriteria of criterion user opinion solicitation using fuzzy linguistic modelling

<table>
<thead>
<tr>
<th>Subcriteria</th>
<th>user₁</th>
<th>user₂</th>
<th>user₃</th>
<th>LOWA operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>User satisfaction</td>
<td>H</td>
<td>VH</td>
<td>VH</td>
<td>VH</td>
</tr>
<tr>
<td>User feedback</td>
<td>L</td>
<td>VL</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>Contact information</td>
<td>VL</td>
<td>VL</td>
<td>H</td>
<td>M</td>
</tr>
</tbody>
</table>

### Table 6. Importance of subcriteria of criterion user opinion solicitation

<table>
<thead>
<tr>
<th>Subcriteria</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>User satisfaction</td>
<td>VH</td>
</tr>
<tr>
<td>User feedback</td>
<td>M</td>
</tr>
<tr>
<td>Contact information</td>
<td>L</td>
</tr>
</tbody>
</table>
5. Discussions and challenges

The importance of evaluating the quality of DLs from the user’s perspective is well recognized by the DL community. However, research on the quality evaluation of DLs based on users’ perceptions seems to be in an early stage. As outlined in Figure 2, there are plenty of definitions for usability and usefulness. Because of this lack of a common lexicon, it is hard to contrast the experimental results obtained by different authors. For instance, Section 3 reports that, according to an experiment conducted by Kani-Zañhibi et al. [36], the criterion supporting collaborative knowledge working has low importance. Nevertheless, the results presented by Blandford et al. [29] support a contrary conclusion.

These 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 [32] uses 41 subjects, and Xie [43] uses 19), the results may be statistically non-significant.

Regarding tools for measuring users’ perceptions, two different approaches have been summarized in Section 4: Likert scales and fuzzy linguistic modelling. However, the two approaches have not been compared in the literature yet. Although Likert scales fall within the ordinal level of measurement, many authors treat individual responses as interval data [18, 32, 43, 45, 50], without any justification for the assumption of interval status for Likert data. As we have noted, if the wrong statistical technique is used, researchers increase the chance of coming to the wrong conclusion about the significance of their research.

To sum up, the user-centred evaluation of DLs requires the following challenges to be tackled:

1. A consensus on standard definitions for usability and usefulness has 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 with fuzzy linguistic modelling have to be identified.

We propose Table 7 as a starting point, which should be developed in future work. According to this table, fuzzy linguistic modelling seems to produce better results (the details are presented in Sections 4.1 and 4.2). On the other hand, Likert scales support measurement of the user’s opinion dispersion.

6. Conclusions

In this paper, we have discussed the state of the art of the quality evaluation of DLs based on users’ perceptions by conducting a structured literature review covering 41 primary studies, and outlining the main advances made up to now. As a result, we have summarized what criteria are being used to evaluate the quality of DLs, the importance of each criterion, and the inter-criteria correlation. We have also provided information about the measures that are derived from these criteria, the measuring instruments used to elicit users’ opinions, 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:

1. the standard definition of usability and usefulness criteria;
2. the minimum necessary requirements to guarantee the validity of experiments on criteria correlation/priorization; and
3. the comparative analysis of existing proposals to obtain evaluations by combining the information collected via measuring instruments.
Acknowledgements

This work was supported by the FEDER funds in FUZZYLING-II Project TIN2010-17876; Andalusian Excellence Projects TIC-05299 and TIC-5991; and ‘Proyecto de Investigación del Plan Propio de Promoción de la Investigación UNED 2011’.

Notes

1 In the literature, *ease of use* and *utility* are sometimes used as synonyms for *usability* and *usefulness*, respectively [27].
2 The purpose of central tendency is to determine the single value that identifies the centre of a distribution and best represents the entire set of scores.

References


