

Testing a Multidimensional and Hierarchical Quality Assessment Model for Digital Libraries

Alexandru Balog

National Institute for Research and Development in Informatics - ICI Bucharest, Romania

e-mail: alexb@ici.ro

Abstract. This paper proposes and tests a multidimensional and hierarchical model for quality evaluation of digital libraries (LibEval). A convenience sample of 252 undergraduate and graduate students was used. Data was collected in October 2010 by means of a paper-and-pencil questionnaire. Results from a confirmatory factor analysis utilizing structural equation modelling techniques confirm the existence of a second-order factor construct (Quality of Digital Library) and five first-order dimensions, namely: interface quality, system quality, information quality, service quality, and contextual factors. The results from this study show empirically that the five dimensions are distinct manifestations of the quality of digital libraries. The user perceives and evaluates the quality of a digital library globally and in each dimension. The LibEval instrument can be utilized to assess the quality of digital libraries from the general user's perspective.

Keywords: digital library, quality model, multidimensional and hierarchical model, model testing, CFA, SEM.

1. Introduction

The term „digital library” (DL) covers different applications and has been used interchangeably for digitized collections, e-journal platforms, network databases, library websites. As Chowdhury [7] pointed out, „a modern digital library is a space – a centre of intellectual activities – with content, available in different forms and formats in a distributed network environment, as well as tools and facilities for user-centric access, use, interactions, collaborations and sharing”.

One major area of library and information science research is the evaluation of digital library. In line with the growing number of operational DLs, there is also growing interest among researchers to investigate the quality of DLs [1], [13], [22], [34], [42], [44]. In the context of Romania, no studies were found addressing this topic. The overall quality of DLs is insufficiently studied and reported [44], and DL quality and evaluation is a very underrepresented research area in the digital library literature [13]. While previous studies have provided useful and interesting information regarding the topic, this study proposes and tests a multidimensional and hierarchical model which can be used to assess the quality of digital libraries. The model is based on Zhang model [44] and DeLone & McLean IS success model [9].

This paper is organized as follows. In Section 2 we briefly review the related work and studies on DL quality. In Section 3 we present the proposed research model, which is followed by

the description of the research methods used and statistical results in Section 4 and 5, respectively. A summary and conclusions are presented in Section 6.

2. Related Works

Researchers and DL practitioners proposed several DL evaluation models, frameworks, and tools, along with various evaluation criteria, indicators, and measures [42]. A way to increase the efficiency of any evaluation method is to provide evaluators with suitable tools able to assist them during the evaluation process [31].

In a meta-analysis study of the evaluation practice, Saracevic [34] has classified the various variables in four main categories, namely: constructs that were evaluated, context in which the evaluation were conducted, criteria that were chosen as a basis for evaluation, and methods that were used during the evaluation. Also, Saracevic [33], [34] suggested six levels of DL evaluation criteria, representing content, technology, interface, service, user, and context. DELOS Network of Excellence has conducted a series of research concerning the evaluation of DLs and developed the DELOS Reference Model [1]. In this model, the notion of „quality” is one of the highest level concepts which groups parameters characterizing the DL behaviour within a given operational domain. Fuhr et al. [12] proposed a scheme for digital library evaluation which contains four dimensions: data/collection, system/ technology, users, and usage. Tsakonias and Papatheodorou [37] further examined the

interactions of DL components. An interaction is composed of three components: the user, the content and the system. They identified three categories of criteria which define relationships among components: usability (user-system), usefulness (user-content), and system performance (content-system).

As noted by Xie [40], within the DL evaluation criteria, usability is the one that has been most investigated. Xie [40], [41] investigated DL evaluation criteria from the users' perspective and classified them in five categories, namely: interface usability, collection quality, service quality, system performance efficiency and user opinion solicitation. Studies offer detailed information about evaluation criteria, their importance in evaluating DL, and relationship between perceived importance of DL evaluation criteria and actual evaluation of DL. Jeng [18] proposed and tested an evaluation model for assessment of the usability of academic digital libraries by examining their effectiveness, efficiency, satisfaction, and easiness in learning. Tsakonas and Papatheodorou [38] explored usefulness and usability of electronic information service systems, such as digital libraries, e-journal platforms, portals, and e-prints. The results indicated that user interaction was affected by both content and system characteristics. Also, ease of use and navigation are the most influential attributes of a system.

The DigiQUAL™ protocol [22] has identified 180 items linked to twelve themes related to DL service quality: design features; accessibility / navigability; interoperability; DL as community for users, developers and reviewers; collection building; role of federations; copyright; resource use; evaluating collections; and sustainability. Gonçalves et al. [13] defined a formal quality model for digital libraries based on the 5S formal framework for digital libraries. Key concepts of a minimal DL (digital object, metadata specification, collection, catalogue, repository, and service) are considered along with corresponding quality dimensions and a set of indicators. Shen [36] formalized the DL integration problem and proposed a model of DL success from the end-user perspective, based on the 5S framework and DeLone-McLean information system success (D&M IS success model) [9].

Recently, Zhang [44] developed a "holistic digital library evaluation model" for various

levels of digital library evaluation according to Saracevic's six levels, in which the most relevant DL evaluation criteria among five groups of stakeholders (administrators, developers, librarians, researchers, users) are identified and organized. The model contains 19 core criteria and 18 group-based criteria. The author constructed the holistic model by using descriptive and inference statistical techniques. The validity of the model was tested through stakeholders' interaction with a real digital library.

More recently, Balog [3], [4] proposed and analyzed a multidimensional and hierarchical model (LibEval) can be used to assess the quality of digital libraries. The model is based on Zhang model [44] and D&M IS success model [9]. It contains 5 dimensions and 19 criteria. The preliminary validity of the model was assessed by using of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

Based on previous research concerning DL evaluation, we believe that additional research is needed in order to examine and make the evaluation criteria operational, as well as to test the validity of the criteria in various settings.

3. Research Model

This study continues the author's previous research [3], [4] on digital library evaluation, and intends to validate the previous results. We have proposed a multidimensional and hierarchical model (LibEval) which synthesizes diverse research in the area of digital library quality models and information systems success.

According to the taxonomy proposed by Law et al., [23], *Quality of digital library* can be viewed as a multidimensional construct. It is a higher level construct underlying its dimensions, with the relationships flowing from the construct to its dimensions, and there is no causal relationship between the dimensions and the construct that represents them. As noted by Edwards [10] this type of construct „may be named *superordinate* because it represents a general concept that is manifested by specific dimensions. The dimensions of a superordinate construct are themselves constructs that function as specific manifestations of a more general construct”.

Multidimensional and hierarchical representation of the DL quality is based primarily on theoretical arguments. Thus, the theory of DL evaluation emphasizes the multidimensional aspects of DL. Cook et al., [8] confirmed that a second-order factor is associated with the provision of high quality in an academic library. Furthermore, in the frameworks developed by Marchionini [26] and Nicholson [29] the "multifaceted" and "perspective" issues suggest that the different dimensions of library quality can be more or less important depending on evaluation objectives, context of use, etc.

The dimensions and the evaluation criteria in the proposed LibEval model are borrowed from existing studies. More specifically, the dimensions are adapted from the DeLone & McLean IS success model [9] and from the levels of the holistic DL evaluation model developed by Zhang [44]. Also, the evaluation criteria are mainly adapted from Zhang [43], [44] and bear the same name as those of Zhang; however, the content is somewhat different from that described by Zhang.

researchers, users) and contains 19 core criteria perceived unanimously as important by all the five groups, and 18 group-based criteria on the top lists of some stakeholder group. Also, the model comprises important evaluation criteria for DL evaluations at six levels (content, technology, interface, service, user, context). A DL evaluation can be conducted flexibly by adopting all core and selective group-based criteria in any of the six levels based on evaluation objectives and target stakeholders' interests".

In LibEval model, the DL quality is defined as the composite relationships of interrelated dimensions consisting of Interface Quality, System Quality, Information Quality, Service Quality, and Contextual Factors. Each of these dimensions has been directly or indirectly investigated in previous studies addressing digital library [5], [15], [28], [32], [36], [42].

The proposed LibEval model is developed from the user's perspective, or, in other words, the identified dimensions and the evaluation criteria are important for the general user. The evaluation criteria are selected from the Zhang

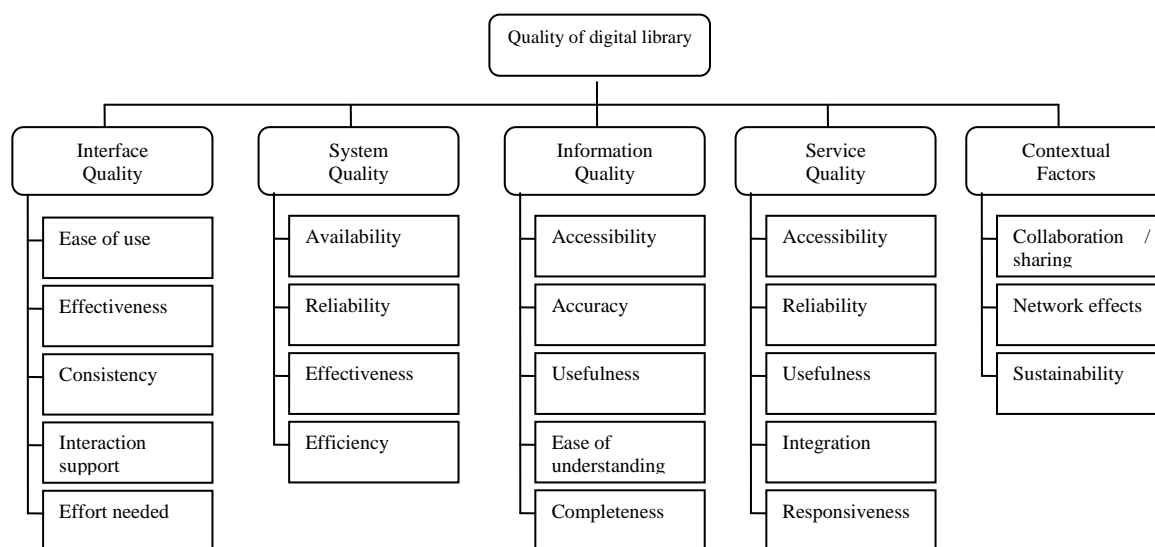


Figure 1. The LibEval research model

The D&M IS success model identifies six interdependent components of IS success: information quality, system quality, service quality, intention to use, user satisfaction, and net benefits. The D&M model has been widely used by researchers for understanding and measuring the dimensions of IS success [30]. As Zhang [44] pointed out, "the holistic digital library evaluation model incorporates diverse viewpoints from different DL stakeholder groups (administrators, developers, librarians,

model, as suggested by author. Therefore, we adopted 14 core criteria from 19, and 7 user-based criteria. Also, we replaced one criterion. The LibEVAL model is presented in Figure 1.

Interface quality refers to user perception of a DL's web site which facilitates and enhances the interactions between the DL and its users. We categorize factors into five attributes: ease of use, effectiveness, consistency, interaction support, and effort needed [44]. Generally, a

highly-quality interface should be intuitive, user-friendly, and should enable users to perform prescribed tasks quickly and without errors. Also, it should provide the search and browsing functions by using different, simple or complex criteria which reduces effort is using a particular library [5]. Navigational aids make it easier for users to follow the logical flow of information and conduct more efficient searches [28], [32]. The visual appearance of a DL should be presented in a consistently manner in terms of layout, colour schemes, font sizes, icons, buttons, etc. Also, terminology is important to a user's capacity to search result in a DL [32], [44].

System Quality refers to user perception of a DL's performance in retrieving and delivering information. We categorize factors into four major attributes: availability, reliability, effectiveness, and efficiency.

The DLs is designed to allow access to the system at remote locations, providing access to content anytime and anywhere, which is critical for promoting the use of DL. Problems with system performance are related to the relevance of the retrieval results and efficiency of the retrieval process [41]. The ability of a system's functions to fit the user's information needs and to reduce the effort of using a DL are important characteristics of the DL [43], [44]. Therefore, the ability of the DL to assist users in performing their tasks better and faster, streamlining work processes and thereby improving their productivity, are critical to the quality of the digital library.

Information Quality refers to user perception of a DL's information which is provided to support the information needs. In this study, the Information Quality is assessed in terms of the degree to which the information is readily accessible, accurate, complete, easy to understand by users, and useful to targeted users in achieving certain goals [19], [24], [39], [44]. Firstly, the information needed for performing a specific task (e.g., finding a conference paper) should be easy to retrieve and to access. Secondly, the information retrieved should be free of errors, complete, and easy of understand. Users want to find the information they need quickly and with little effort. It is, therefore, important to have only the most relevant information which meets the information needs.

Service Quality refers to user perception of a DL service (e.g., online access catalogue, e-resources such as online database, e-books and e-journals), provided by the digital library through web site to support the information needs and individual user's specific tasks [15], [44]. Service quality is important in the context of digital library, because digital communication lacks face-to-face contacts. Therefore, DL service should be accessible and reliable, should provide effective online support, should be able to provide users with positive and prompt responses, and should be useful to users in helping them to achieve certain goals [43].

Contextual factors refer to user perception of a DL's characteristics which facilitate the individual user's task implementation. According to Zhang [43], evaluation at the context level assesses how well a given DL fits into larger contextual practices (e.g., institutional, social, cultural, economic, legal). In this study, we selected the following criteria from the Zhang model: collaboration /sharing, network effects, and sustainability. A DL should have processes and products which are shared among stakeholders and/or institutions. Also, a DL should have lots of incoming and outgoing links to other web resources, and a given organization should have adequate resources to sustain a digital library.

4. Research Methods

This study is part of a larger research project which aimed to develop and validate the LibEval instrument to evaluate the quality of digital libraries. Based on the studies of Koufteros [20], Segars [35], and Hair et al. [14], the research methodology used in the project included four phases: instrument development, exploratory study, confirmatory study, and testing of the structural model.

The purpose of the first phase was to define the conceptual domain and dimensionality of the construct „Quality of digital library”, as well as to generate the scale items. A comprehensive literature review was used to develop a working definition of DL quality, to delineate its dimensions, and to generate a set of items for measuring these dimensions. We designed a questionnaire to investigate user's perceptions with regard to digital library quality (22 items),

their satisfaction (3 items) and their continued usage intention (3 items).

The second phase focused on identifying the underlying dimensions of the LibEval scale, examining dimensionality and establishing the reliability and the validity of the dimensions. Using the questionnaire developed in the first phase and data collected from three universities from Romania, an EFA was conducted. Then, CFA was used for the validation of key properties of scales, including unidimensionality, reliability and validity. In the iterative process of purifying the scales three items were excluded from further analysis. In the final version of the model, 19 items were used to measure the DL quality. The first phase and the second phase are well documented in [3], [4].

The third phase sought to confirm the structure of the LibEval scale by analyzing new data obtained from another survey. We analysed several measurement models and selected one model as the best representing measurement model. The central focus of the paper is on this phase.

In the fourth phase a structural model was tested in a nomological network where the constructs „satisfaction” and „continued usage intention” are hypothesized to be caused by the construct „Quality of digital library”. They are not the focus of this paper and thus no further discussion of those matters is pursued in the paper.

We describe below the methods that were used in the third phase (confirmatory study).

4.1 Data collection and sample

A survey was conducted for data collection by using the questionnaire previously developed. The questionnaire was developed in Romanian. All items were phrased positively and were measured on a 7-point Likert-type scale, ranging from 1 (strongly disagree) to 7 (strongly agree). The items measuring quality of DL are detailed in Appendix 1.

The population for this study was composed of students from one of the Romanian public university. Out of 300 questionnaires distributed, 259 were returned. After dropping responses from incomplete questionnaires and those with excessive missing data, 252 usable responses were obtained, producing a response rate of 84%. Since a minimum sample size of

150, 200 or more was suggested for CFA [2], [14], the sample size was considered acceptable.

Among the 252 respondents, 151 (59.9%) of them were male and 101 (40.1%) were female. The age profile of the respondents ranged from 20 years to 40 years with most of the respondents (42.1 %) lying between 21 years and 25 years. The majority of respondents (68.3 %) had experience in the use of digital libraries.

4.2 Analytical procedures

In order to assess the underlying factor structure in the proposed scale [4], a CFA using structural equation modelling (SEM) approach was conducted. Testing of the measurement instrument was done by the paradigm proposed by Koufteros et al., [21] for assessing second-order measurement models and by the set of recommendations elaborated by MacKenzie et al., [25] for developing valid measures.

According to [21], the testing process includes the construction and the assessment of several alternative measurement models. To compare the measurement models and assess whether a second-order is plausible, various fit indices can be compared. Based on an analysis of relevant research, we selected indices and cut-off values according to recommendations from Hu & Bentler [16] and Hair et al., [14]. The following goodness-of-fit measures were used in this study: normed chi-square (χ^2/df), Tucker-Lewis index (TLI), comparative fit index (CFI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA). In summary, to support model fit it is desirable to exhibit: the normed chi-square should be less than 3, TLI and CFI should exceed 0.95, SRMR should be less than 0.05, and RMSEA should be less than 0.05 and its 90% confidence interval to be not excessively wide.

The next step was to assess the convergent and discriminant validity. Convergent validity can be assessed by examining the loading and their statistical significance through *t*-values, the item reliability, the construct reliability, and the average variance extracted [11]. Factor loadings of the indicators are estimates of the validity of those variables to its underlying construct [20]. To be considered adequate, the factor loadings of all standardized items should be greater than 0.70 and *t*-values greater than $|1.96|$ at the 0.05 level. Item reliability

indicates the amount of variance in an item due to the underlying construct rather than error and it should be greater than 0.50 [20]. The composite reliability (CR) measures the internal consistency on the indicators measuring given construct and should be at least 0.70. The average variance extracted (AVE) measures the amount of variance that is captured by the construct in relation to the amount of variance due to measurement error and it should be greater than 0.50 [11].

Discriminant validity testing is aimed at examining the extent to which constructs are distinct from each other. We used two approaches in evaluating discriminant validity. Firstly, discriminant validity is demonstrated when the AVE for each construct is higher than the squared correlation between any two constructs [11]. Secondly, discriminant validity can be assessed by constraining the correlation between any two constructs to 1.0 and then performing a χ^2 difference test between the constrained and unconstrained models. The test should be performed for one pair of constructs at a time. When a χ^2 value for the constrained model is significantly greater than that of the unconstrained model, discriminant validity is achieved [2]. The χ^2 difference between the constrained and unconstrained models always has one degree of freedom; therefore, the χ^2 difference should be greater than 3.84 to be considered significant.

5. Analysis and Results

5.1 Data screening

Data analysis was carried out using the SPSS 16.0 for Windows software. The descriptive statistics for each item are shown in Table 1.

All mean scores are greater than 4.0 (i.e., midpoint on the seven-point Likert scale), with a range of 4.75–5.26, indicating that variables were considered important when assessing a digital library. Of the 19 variables in the scale, the “effort needed” (SIT5) item had the highest mean score (M=5.26; SD=1.32) and the “reliability” (SRV2) item had the lowest mean score (M=4.75; SD=1.25). The standard deviations range from 1.14 to 1.45 indicated that students responded consistently to the instrument as a whole and to its five subscales.

Prior to analyses, data was examined for the presence of univariate and multivariate outliers

[14], which might attenuate the results. The former was analyzed through standardized scores ($|z| \geq 3.30$) and the latter through Mahalanobis distance ($p < .001$). No univariate outliers were identified and no multivariate outliers were found.

Table 1. Descriptive statistics of the items (N=252)

Construct	Item	Mean	SD	Skewness	Kurtosis	
Interface	SIT1 ^{*)}	5.00	1.37	-.111	-.489	
	Quality	SIT2	5.08	1.33	-.426	-.150
		SIT3	5.02	1.37	-.140	-.400
		SIT4	5.01	1.31	-.193	-.506
		SIT5	5.26	1.32	-.180	-.751
System	SYS1	5.00	1.31	-.344	-.094	
	Quality	SYS2	5.23	1.16	-.062	-.409
		SYS3	5.23	1.17	-.352	-.167
		SYS4	5.24	1.16	-.245	-.031
Information	INF1	5.06	1.17	.122	-.478	
	Quality	INF2	5.14	1.14	-.062	-.384
		INF3	4.96	1.21	-.067	-.400
		INF4	5.13	1.19	-.164	-.259
Service	SRV1	4.96	1.28	-.330	-.144	
	Quality	SRV2	4.75	1.25	-.014	-.075
		SRV3	5.06	1.45	-.315	-.538
Context	CXT1	4.87	1.26	-.184	-.044	
	CXT2	4.84	1.38	-.033	-.465	
	CXT3	5.02	1.29	-.298	-.087	

^{*)} The abbreviation refers to the questions shown in Appendix 1

Also the normality of each of the 19 items was investigated in terms of its skewness (-.426 to .122) and kurtosis (-.751 to -.031). These values were all within the recommended level, supporting the moderate departure from normality for all variables. For this reason, no cases were deleted.

5.2 Evaluating the goodness-of-fit of the measurement models

In the present study there were constructed four alternative measurement models. The first model is hypothesized to include one global factor with 19 reflective indicators. Model M2 hypothesizes five first-order uncorrelated factors. Model M3 (Figure 2) hypothesizes five first-order correlated factors. Finally, the model

M4 (Figure 3) includes one second-order construct and five first-order factors with corresponding indicators. The measurement models were analyzed using the AMOS 7.0 software using a covariance matrix as input and maximum likelihood estimation method. The goodness of fit indices of the measurement models are shown in Table 2.

The results for Model M1 show a range of fit results that are indicative for a poor model fit. The results for Model M2 also reflect a range of fit results that fall outside the acceptable ranges. While the fit indices, except SRMR, were considerably improved over those from Model M1, they failed to achieve acceptable levels for any of the indices selected for this study. This result was anticipated, since the Model M2 proposes five independent factors, all related to digital library quality, without any interactions between them.

The Model M3 (Figure 2), hypothesizes that: (a) responses to the 19-items scale can be explained by five factors; (b) each item has a nonzero loading on the factor that it was designed to measure and a zero loading on all other factors; (c) the five factors, consistent with the theory, are correlated; (d) measurement error terms are uncorrelated [6]. The Model M3 was designed to test the relationships between five factors and their 19 observed indicators. The number of observed variances and covariances ($19[19+1]/2$) was 190 data points, and the number of estimated parameter in the model was 48 (14 regression weights, 10 covariances, and 24 variances). The five correlated first-order factors model was overidentified (the number of observed variances and covariances is greater than the number of estimated parameter), and tested with 142 degrees of freedom (190-48).

The results for Model M3 (Table 2) indicate a good level of fit of the proposed model with sample data. Although the χ^2 test is significant ($\chi^2=203.268$, $df=142$, $p<0.01$), other fit indices indicated a good fit with $\chi^2/df=1.431$, TLI=0.972, CFI= 0.977, SRMR=0.040, RMSEA =0.041 and its 90% confidence interval is fairly narrow (i.e., 0.028-0.054). However, it is possible that the multivariate normality might exist as Mardia's multivariate coefficient was not acceptable (33.933), and this could result in the estimate of χ^2 statistic of the model being distorted. Thus, the Bollen-Stine bootstrapping was used to redistribute the

data for overcoming the problem of multivariate non-normality. The Bollen-Stine p was proved satisfactory, ($p=.069$), which was not significant at $p < .05$.

The examination of the standardized residual values of the covariance matrix has shown that they do not exceed the value of |1.96|, and there is no need for respecification. Also, the inspection of the modification indices and the completely standard expected change has shown that it does not appear to be a reason for respecification and does not justify an alternative specification [19]. The correlations between first-order factors are moderate, with values ranging from 0.46 to 0.63. As argued by Koufteros et al. [21] in instances where first-order factors exhibit moderate correlations, a measurement model specification such as the one represented by Model M3 would be advisable.

The Model M4 (Figure 3), hypothesizes a priori that: (a) responses to the 19-items scale can be explained by five first-order factors, and one second-order factor; (b) each item has a nonzero loading on the first-order factor that it was designed to measure, and a zero loading on the other for first-order factors; (c) measurement error terms associated with each item are uncorrelated; (d) covariation among the five first-order factors is explained fully by their regression on the second-order factor [6]. The Model M4 was designed to test the relationships between five first-order factors and one second-order factor. The number of observed variances and covariances ($19[19+1]/2$) was 190 data points, and the number of estimated parameter in the model was 43 (19 regression weights and 24 variances). The second-order model was overidentified, and tested with 147 degrees of freedom (190-43). In addition, Byrne [2001] suggests that with a second-order model, it is necessary to "check the identification status of the higher order portion of the model". Given that there are 15 data points ($5[5+1]/2$) and the number of estimated parameter was 10 (five factor loading and five residuals), the model was overidentified with five degree of freedom.

The results for Model M4 (Table 2) indicate a good level of fit of the proposed model with sample data. Although the χ^2 test is significant ($\chi^2=206.843$, $df=147$, $p<0.01$), other fit indices indicated a good fit with to the data: $\chi^2/df=1.407$, TLI=0.974, CFI=0.978,

SRMR=0.043. Also, the RMSEA value is 0.040, with the 90% confidence interval ranging from 0.026 to 0.053. The upper bound of 90% CI is less than the cut-off value 0.06

suggested by Hu & Bentler [16], and the p value for the test of closeness of fit is equal to .900. These results provide adequate support for the existence of the second-order structure.

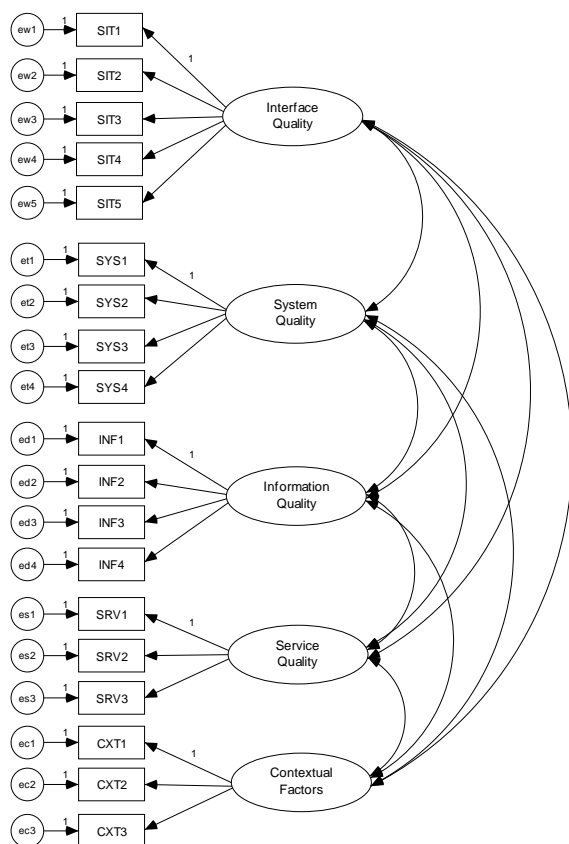


Figure 2. M3: five first-order correlated factors

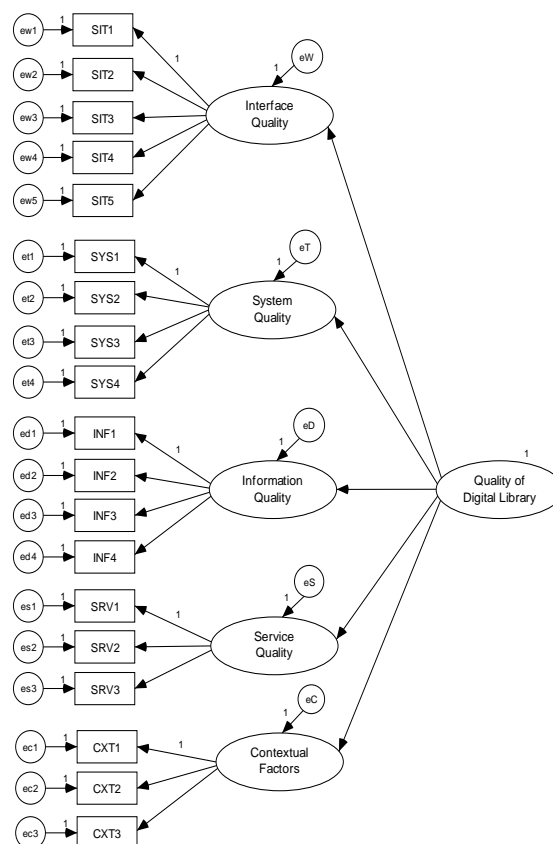


Figure 3. M4: five first-order factors, one second-order factor

Table 2. Goodness of fit indices for alternative measurement models (N=252)

Goodness of fit indices	Measurement models				
	Recommended value ¹	Model M1 One first-order factor	Model M2 Five first-order factors - uncorrelated	Model M3 Five first-order factors - correlated	Model M4 Five first-order factors, one second-order factor
χ^2	$\leq 3df$	1161.374	541.485	203.268	206.843
df		152	152	142	147
p value	$p > .05$.000	.000	.001	.001
χ^2 / df	≤ 3	7.641	3.562	1.431	1.407
TLI	$> .95$.577	.837	.972	.974
CFI	$> .95$.624	.855	.977	.978
SRMR	$\leq .05$.111	.289	.040	.042
RMSEA,	$\leq .05$ or $\leq .06$.163	.101	.041	.040
p value,	$pclose > .05$.000	.000	.866	.900
90% CI	close to RMSEA	.154-.171	.092-.110	.028-.054	.026-.053

¹ adapted from Byrne [6], Hu & Bentler [16] and Hair et al. [14]. χ^2 : chi-squared; df : degree of freedom; TLI: Tucker-Lewis Index; CFI: Comparative Fit Index; SRMR: Standardized Root Mean Square Residual, RMSEA: Root Mean Square Error of Approximation; 90% CI: 90 Percent Confidence Interval for RMSEA.

As expected, the values of fit indices for Model M3 and Model M4 are similar, with the exception of the chi-square statistic. The second-order factor is explaining the covariation among first-order factors in a more parsimonious way. According to Marsh & Hocevar [27], even when the higher-order model is able to explain effectively the factor covariations, the goodness-of-fit of the higher-order model can never be better than the corresponding first-order model. Also, if a hierarchical model fits the data nearly as well as a corresponding first-order model, then there is support for the second-order model. As Koufteros et al., [21] pointed out, a second-order model that rivals the performance of a first-order correlated model can be an attractive alternative.

In addition to the evaluation of the goodness of fit statistics, we also assessed the capacity of the second-order model to explain the covariation among the first-order factors. Target (T) coefficient, as suggested by Marsh & Hocevar [27], defined as the ratio of the chi-square values for the first-order model and the second-order model, was computed. Target coefficient has an upper bound of 1.0. A target coefficient higher than 0.90 may indicate the existence of a second-order factor construct since most of the variation shared by the first-order factors is explained by the single second-order factor.

The calculated target coefficient between the first-order model and second-order model is 0.98 ($T=203.268/206.843$) supporting that the second-order factor explains almost all of the covariation among first-order factors. The second-order model represents a more parsimonious representation of observed covariances (five paths in contrast to ten correlations) and it should be accepted over the first-order model as a „truer” representation of model structure [27].

Therefore, for Model M3 and Model M4 all the goodness-fit indices exceeded their respective common acceptance, thus demonstrating that the measurement models exhibited a fairly good fit with the data collected. We could proceed to evaluate the psychometric properties of the measurement models in terms of reliability, convergent validity, and discriminant validity.

5.3 First-order measurement model

The results provided support for convergent validity (Table 3). All standardized factor loadings were statistically significant (t -values > 1.96), and ranged from 0.715 to 0.859, which were above the recommended threshold of 0.70. The item reliability (R^2) values are above the suggested standard of 0.5 [20].

As shown in Table 3, the composite reliability (CR) of each construct is above the minimum level of 0.70, ranging from 0.815 to 0.896. This suggests that the items are sufficiently representative on their respective construct.

Table 3. Results of convergent validity

Construct / items	Std. loading (λ)	t-value (for λ)	R ²	CR	AVE
Interface quality (SIT)				0.896	0.634
SIT1	0,841	- ^a	0,708		
SIT2	0,759	13,73	0,575		
SIT3	0,828	15,57	0,685		
SIT4	0,800	14,82	0,640		
SIT5	0,749	13,49	0,561		
System quality (SYS)				0.872	0.630
SYS1	0,803	- ^a	0,646		
SYS2	0,859	14,61	0,738		
SYS3	0,792	13,38	0,627		
SYS4	0,715	11,81	0,511		
Information quality (INF)				0.864	0.615
INF1	0,784	- ^a	0,615		
INF2	0,828	13,42	0,685		
INF3	0,776	12,53	0,602		
INF4	0,746	12,00	0,557		
Service quality (SRV)				0.882	0.713
SRV1	0,855	- ^a	0,731		
SRV2	0,829	15,24	0,687		
SRV3	0,849	15,65	0,720		
Contextual factors (CXT)				0.815	0.596
CXT1	0,794	- ^a	0,630		
CXT2	0,773	11,29	0,597		
CXT3	0,748	11,04	0,560		

^a Indicates a parameter fixed at 1.0 in the original solution

The values of average variance extracted (AVE) are all above the minimum recommended level of 0.50, ranging from 0.596 to 0.713. Among the AVEs of the constructs, Service Quality had the highest value of 0.713, indicating that 71.3% of the variance in the specified indicators was accounted for by the construct. These results suggest the strong relationships between respective items and constructs and thus provide evidence of convergent validity.

The discriminant validity was evaluated through two methods. Using the squared correlations test [11], the results in Table 4 show that all squared correlations between two constructs were found to be smaller than the AVE measures of both constructs. For example, in Table 5, the lowest AVE value is 0.596, and the highest value of the square of correlation coefficient is 0.394.

Table 4. Squared correlations and AVEs

	SIT	SYS	INF	SRV	CXT
SIT	0.634				
SYS	0.260	0.630			
INF	0.394	0.235	0.615		
SRV	0.325	0.228	0.226	0.713	
CXT	0.256	0.215	0.215	0.224	0.596

Note: Diagonal elements are the AVE for each of the five constructs. Lower off-diagonal elements are the squared correlations between constructs

These results suggest that the items share more common variance with their respective constructs than any variance the construct shares with other constructs and thus provide evidence of discriminant validity.

Using the second method, testing discriminant validity through pair wise χ^2 difference tests require the estimation of 10 covariance structures and evaluation of ten χ^2 difference. Thus, to evaluate all 5 constructs, 10 separate constrained models were conducted and compared with their corresponding unconstrained model. The chi-square differences of the 10 pairs of the models were significant at $p < .001$ with values greater than 3.84, ranging from 4.68 to 26.43. This provides evidence of discriminant validity.

5.4 Second-order measurement model

For second-order constructs with reflective dimensions, the convergent validity of the indicators of the first-order factors (dimensions) can be assessed as previously described for first-order measurement model. Estimates of item validity and reliability are not sensitive to the addition of a second-order factor. Conclusions concerning the validity and reliability of the nineteen items would be the same. Model M4 has the additional advantage of providing estimates of the validity and reliability of the dimensions.

Further empirical support for convergent validity and for acceptance of the second-order factor model is found in magnitude and significance of estimated parameters as well as the amount of variance explained [2]. As noted by MacKenzie et al. [25], for second-order constructs with reflective indicators it is also important to examine the convergent validity of the set of first-order factors as reflective indicators of the second-order construct.

The validity of each dimension can be tested by examining whether it is significantly related to the second-order construct. All standardized factor loadings of the first-order factors on the second-order factor are large and exhibit high t -values at the 0.001 significance level. Specifically, the results indicated that Interface Quality ($\gamma = 0.812$, t -value=11.55) was the strongest dimension of the second-order factor (Quality of Digital Library), followed by Information Quality ($\gamma = 0.737$, t -value=9.78), Service Quality ($\gamma = 0.695$, t -value=9.82), System Quality ($\gamma = 0.660$, t -value = 8.97), and Contextual factors ($\gamma = 0.650$, t -value=8.44). The results suggest that the users evaluate the DL quality on the confirmed five dimensions, but they also viewed the DL quality as a higher-order factor that captured a meaning common to all the dimensions.

The construct reliability (CR) for the second-order construct is 0.837, above the minimum recommended level of 0.70. This suggests that the dimensions are sufficiently representative on the second-order construct. The value of average variance extracted (AVE) for second-order construct is 0.509, above the minimum level of 0.50. These results suggest that, on average, more than half of the variance in the first-order dimensions is shared with second-order construct and thus provide evidence of convergent validity.

The correlation coefficients of the five dimensions ranged from 0.46 to 0.63, and were below the recommended threshold of 0.85. Also, using the squared correlations test [11], the results have shown that all squared correlations between two constructs were found to be smaller than the AVE measures of both constructs. These results provide evidence of discriminant validity.

Therefore, based on the above results, we concluded that the second-order factor structure was well supported. This suggested that students evaluated digital library quality not only on the basis of the five dimensions but they also evaluated digital library quality as a whole.

6. Conclusions

This study presented an empirically validated model for measuring DL quality. The 19-item LibEval instrument was demonstrated to produce acceptable reliability estimates, and the empirical evidence supported its convergent validity and discriminant validity.

The study contributes to the existing literature on library quality by studying the construct and its dimensions in Romanian context. The dimensions are adapted from the D&M IS success model [9] and from the levels of the holistic DL evaluation model developed by Zhang [44]. The items obtained in this study bear the same name as those of Zhang; however, the content is somewhat different from that described by Zhang.

Multidimensional and hierarchical representation of the quality of digital libraries is based primarily on theoretical arguments. The results from this study showed empirically that the five dimensions are distinct manifestations of the quality of digital libraries. The user perceives and evaluates the quality of digital library globally and in each dimension. Interface Quality was the strongest dimension of the Quality of Digital Library, followed by Information Quality, Service Quality, System Quality, and Contextual Factors.

Although we cannot claim to have fully captured the dimensions of DL quality, it may be argued that we are close to it because the second-order factor extracts the underlying commonality among dimensions. Specifying quality DL as a second order construct makes it possible not only to measure and examine the

importance of quality for the user, but also to highlight and compare the relative importance of each dimension.

The results emphasize the importance of assuming a multidimensional approach and encourage library managers to include the measures of Interface Quality, System Quality, Information Quality, Service Quality, and Contextual Factors into their present evaluation techniques of digital library quality. The LibEval instrument can be used to assess the quality of digital libraries from the general user's perspective. Besides making an overall assessment, the LibEval instrument can be used to compare quality for different digital libraries with specific factors (dimensions). If a library finds itself lacking in any of these dimensions, then it may do a more detailed analysis and take the necessary corrective actions.

There are inherent limitations of this research that should suggest some research directions. Firstly, while this study has been conducted with methodological rigor, the findings should be used with caution. Since this study only examines the quality of a digital library using a convenience sample from Romania, the results may not be generalised to other systems and cultural areas. Additional research efforts should be conducted in order to validate the proposed model and findings in other setting. Secondly, the second-order construct conceptualized in this study has been called a superordinate construct by Edwards [10], and a Type I construct (reflective first-order, reflective second-order) by Jarvis et al. [17]. Thus, a line of research could be the conceptualization of a model based on the criteria and recommendations for the development of formative models [17]. Thirdly, the nomological validity should be validated using structural equation modelling (SEM) in the future. Thus, the nomological validity of the LibEval instrument should be validated through investigating the causality between the construct and its theoretically related antecedents or consequences. Demonstrating the validity of the model can have important implications in terms of policy and development practices and improve the quality of digital libraries and providing library services.

Acknowledgements

This work was supported by the LibEval research project funded by ANCS, contract no. 11012/2007.

Appendix 1. List of items

Code	Items	Description
Interface Quality		
SIT1	ease of use	Interface of the DL is easy of use, intuitive, and friendly
SIT2	effectiveness	Interface of the DL provides functions which support easily finding of the information needed
SIT3	consistency	Interface of the DL is presented consistently regarding terminology, colour, size, and fonts.
SIT4	interaction support	The interface provides search and browsing functions user-friendly and performant by using various criteria
SIT5	effort needed	The interface is designed in a way that no require user's extra effort in order to interact with the library / to find desired information
System Quality		
SYS1	availability	The DL is available and accessible permanently.
SYS2	reliability	The DL functions technically free errors.
SYS3	effectiveness	The DL has capability to respond with relevant elements to my information needs.
SYS4	efficiency	The DL operates in a time saving and effort saving manner
Information Quality		
INF1	accessibility	The information needed for performing a specific task is easily retrievable and accessible.
INF2	accuracy	The information provided by the digital library is correct, reliable, and free of errors.
INF3	usefulness	The information provided by the DL is relevant, applicable and support the information needs for the task at hand
INF4	ease of understanding	The information provided by the DL is easy to understand, terminology is clear and consistent
INF5 ^{*)}	completeness	The information provided by the DL is complete and volume of available information is appropriate for my information needs
Service Quality		
SRV1	accessibility	The services provided by the DL are accessible.
SRV2	reliability	The services provided by the DL are correct, trusted and meet the information needs.
SRV3	usefulness	The services provided by the DL are helpful for me in achieving goals
SRV4 ^{*)}	integration	The services provided by the DL are integrated into the main flow of DL use
SRV5 ^{*)}	responsiveness	The services provided by the DL meet needs regarding quick and prompt responses.
Contextual factors		
CXT1	collaboration / sharing	The DL is sharing resources / services so that allows collaboration with other libraries and users
CXT2	network effects	The DL is linked to other web resource through incoming and/or outgoing links
CXT3	sustainability	I have perceived that the institution is capable to develop and to maintain the DL in financial and technical aspects.

^{*)} items eliminated in the previous exploratory study [4], and unused in this paper (the confirmatory study)

REFERENCES

1. AGOSTI, M., N. FERRO, E. A. FOX, M. A. GONCALVES, **Towards a Reference Quality Model for Digital Libraries**, First Intl. Workshop on DLs Foundations, Vancouver, Canada, June 23, 2007.
2. ANDERSON, J. C., D. W. GERBING, **Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach**. Psychological Bulletin, vol. 103(3), 1998, pp. 411-423.
3. BALOG, AL., **Dezvoltarea preliminară a modelului LibEval de evaluare a calității**

- bibliotecilor digitale.** Revista Română de Informatică și Automatică, vol. 21(1), 2011, pp. 11-26.
4. BALOG, AL., **Dezvoltarea și testarea unui model ierarhic de evaluare a calității bibliotecilor on-line: un studiu exploratoriu.** Revista Română de Interacțiune Om-Calculator nr. 4, 2011, pp. 1-22.
 5. BERTOT, J. C., J. T. SNEAD, P. T. JAEGER, C. R. MCCLURE, **Functionality, Usability, and Accessibility. Iterative User-centered Evaluation Strategies for Digital Libraries.** Performance Measurement and Metrics 7 (1), 2006, pp. 17-28.
 6. BYRNE, B., **Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming.** Lawrence Erlbaum Association, Publishers, 2001.
 7. CHOWDHURY, G., **From Digital Libraries to Digital Preservation Research: the Importance of Users and Context.** J. of Documentation 66(2), 2010, pp. 207-223.
 8. COOK, C., F. HEATH, B. THOMPSON, **Users' Hierarchical Perspectives on Library Service Quality: A LibQUAL+ Study.** College & Research Library, 2001, pp. 147-153.
 9. DELONE, W. H., E. R. MCLEAN, **The DeLone and McLean Model of Information Systems Success: A Ten-year Update.** Journal of MIS vol. 19(4), 2003, pp. 9-30.
 10. EDWARDS, J. R., **Multidimensional Constructs in Organizational Behavior Research: An Integrative Analytical Framework.** Organizational Research Methods vol. 4(2), 2001, pp. 144-192.
 11. FORNELL, C., D. F. LARCKER, **Evaluating Structural Equations Models with Unobservable Variables and Measurement Error.** Journal of Marketing Research vol. 18(1), 1981, pp. 39-50.
 12. FUHR, N., G. TSAKONAS ET AL., **Evaluation of Digital Libraries.** Intl. J. on Digital Libraries, vol. 8(1), 2007, pp. 21-38.
 13. GONÇALVES, M. A., B. L. MOREIRA, E. A. FOX, **What is a Good Digital Library? A Quality Model for Digital Libraries.** Inf. Proc. & Manag., 43, 2007, pp. 1416-1437.
 14. HAIR, J. F., W. C. BLACK, B. J. BABIN, R. E. ANDERSON, R. L. TATHAM, **Multivariate Data Analysis.** 6th ed., Prentice Hall, 2006.
 15. HERNON, P., P. CALVERT, **E-service Quality in Libraries: Exploring its Features and Dimensions.** Library & Info. Science Research. 27(3), 2005, pp.377-404.
 16. HU, L. T., P. M. BENTLER, **Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives.** Structural Equation Modeling vol. 6(1), 1999, pp. 1-55.
 17. JARVIS, C. B., S B. MACKENZIE, P. M. PODSAKOFF, **A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research.** J. of Consumer Research vol. 30(2), pp. 199-218.
 18. JENG, J., **Usability Assessment of Academic Digital Libraries: Effectiveness, Efficiency, Satisfaction, and Learnability.** Libri, vol. 55(2-3), 2005, pp. 96-121.
 19. KNIGHT, S. A., J. M. BURN, **Developing a Framework for Assessment Information Quality on the World Wide Web.** Informing Science nr. 8, 2005, pp. 159-172.
 20. KOUFTEROS, X. A., **Testing a Model of Pull Production: a Paradigm for Manufacturing Research using Structural Equation Modeling.** Journal of Operations Management, vol. 17(4), 1999, pp. 467-488.
 21. KOUFTEROS, X. A., S. BABBAR, M. KAIGHOBADI, **A Paradigm for Examining Second-order Factor Models Employing Structural Equation Modeling.** Intl. J. of Prod. Ec. 120, 2, 2009, pp. 633-652.
 22. KYRILLIDOU, M., S. GIERSCH, **Developing the DigiQUAL Protocol for Digital Library Evaluation,** in: JCDL'05: Proc. the 5th ACM/IEEE-CS Joint Conf. on DLs, NY, ACM Press, 2005, pp. 172-173.
 23. LAW, K. S., C. S. WONG, W. H. MOBLEY, **Toward a Taxonomy of Multidimensional Constructs,** Academy of Management Review vol. 23(4), 1998, pp. 741-755.
 24. LEE, Y. W., D. M. STRONG, B. K. KAHN, R. Y. WANG, **AIMQ: a Methodology for Information Quality Assessment.** Information & Management 40, 2, 2002, pp. 133-146.

25. MACKENZIE, S. B., P. M. PODSAKOFF, N. P. PODSAKOFF, **Construct Measurement and Validation Procedures in MIS and Behavioral Research: Integrating New and Existing Techniques.** MIS Quarterly, vol. 35(2), 2011, pp. 293-334.
26. MARCHIONINI, G., **Evaluating Digital Libraries: A Longitudinal & Multifaceted View.** Library Trends, vol. 49(2), 2000, pp. 304-333.
27. MARSH, H. W., D. HOCEVAR, **Application of Confirmatory Factor Analysis to the Study of Self-concept: First- and Higher Order Factor Models and Their Invariance Across Groups.** Psych. Bulletin, vol. 97(3), 1985, pp. 562-582.
28. MASREK, M. N., A. JAMALUDIN, S. A. MUKHTAR, **Evaluating Academic Library Portal Effectiveness. A Malaysian case study.** Library Review 59, 3, 2010, pp. 198-212.
29. NICHOLSON, S., **A Conceptual Framework for the Holistic Measurement and Cumulative Evaluation of Library Services.** Journal of Documentation vol. 60,(2), 2004, pp. 164-182.
30. PETTER, S., W. DELONE, E. MCLEAN, **Measuring Information Systems Success: Models, Dimensions, Measures, and Interrelationships,** European J. of Info. Sys. vol. 17(3), 2008, pp. 236-263.
31. PRIBEANU, C., **A Usability Assistant for the Heuristic Evaluation of Interactive Systems,** Studies in Informatics and Control vol. 18(4), 2009, pp. 355-362.
32. RAMAYAH, T., **Interface Characteristics, Perceived Ease of Use and Intention to Use an Online Library in Malaysia.** Information Development, vol. 22(2), 2006, pp. 123-133.
33. SARACEVIC, T., **Digital Library Evaluation: Toward an Evolution of Concepts.** Library Trends vol. 49(3), 2000, pp. 350-369.
34. SARACEVIC, T., **Evaluation of Digital Libraries: An Overview,** in: Agosti, M., Fuhr, N. (eds.): Notes of the DELOS WP7 Workshop on the Evaluation of DLs. Padova, 2004.
35. SEGARS, A. H., V. GROVER, **Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement,** MIS Quarterly, vol. 22(2), 1998, pp. 139-163.
36. SHEN, R., **Applying the 5S Framework to Integrating Digital Libraries.** PhD thesis, Virginia Tech CS Department, Virginia, 2006. <http://scholar.lib.vt.edu/theses/available/etd-04212006-135018/>.
37. TSAKONAS, G., C. PAPTAEODOROU, **Exploring Usefulness and Usability in the Evaluation of Open Access Digital Libraries,** Information Processing and Management, vol. 44(3), 2008, pp. 1234-1250.
38. TSAKONAS, G., C. PAPTAEODOROU, **Analysing and Evaluating Usefulness and Usability in Electronic Information Services.** J. of Information Sci., vol. 32(5), 2006, pp. 400-419.
39. WANG, R. Y., D. M. STRONG, **Beyond Accuracy: What Data Quality Means to Data Consumers,** Journal of MIS, vol. 12(4), 1996, pp. 5-34.
40. XIE, H., **Evaluation of Digital Libraries: Criteria and Problems from Users' Perspectives.** Library & Information Science Research, vol. 28(3), 2006, pp. 433-452.
41. XIE, H., **Users' Evaluation of Digital Libraries (DLs): Their Uses, Their Criteria, and Their Assessment.** Inf. Processing and Management. vol. 44(3), 2008, pp. 1346-1373.
42. XIE, I., S. JOO, **Evaluation Constructs and Criteria for Digital Libraries: a Document Analysis.** TEFKO 2010, Nov. 7-8, 2010, Brunswick, NJ, USA.
43. ZHANG, Y., **Developing a Holistic Model for Digital Library Evaluation.** Dissertation, May 2007. <http://hdl.rutgers.edu/1782.2/rucore10001600001.ETD.17104> [accessed 20 June 2010].
44. ZHANG, Y., **Developing a Holistic Model for Digital Library Evaluation.** Journal of the American Society for Information Science and Technology vol. 61(1), 2010, pp. 88-110.