Evaluating a digital humanities research environment: the CULTURA approach

Christina M. Steiner, Maristella Agosti, Mark S. Sweetnam, Eva-C. Hillemann, Nicola Orio, Chiara Ponchia, Cormac Hampson, et al.

International Journal on Digital Libraries

ISSN 1432-5012

Int J Digit Libr
DOI 10.1007/s00799-014-0127-x

ONLINE FIRST

Springer

Int J Digit Libr
DOI 10.1007/s00799-014-0127-x

Evaluating a digital humanities research environment: the CULTURA approach

Christina M. Steiner · Maristella Agosti · Mark S. Sweetnam · Eva-C. Hillemann · Nicola Orio · Chiara Ponchia · Cormac Hampson · Gary Munnelly · Alexander Nussbaumer · Dietrich Albert · Owen Conlan

Received: 23 December 2013 / Revised: 30 August 2014 / Accepted: 2 September 2014

Abstract Digital humanities initiatives play an important role in making cultural heritage collections accessible to the global community of researchers and general public for the first time. Further work is needed to provide useful and usable tools to support users in working with those digital contents in virtual environments. The CULTURA project has developed a corpus agnostic research environment integrating innovative services that guide, assist and empower a broad spectrum of users in their interaction with cultural artefacts. This article presents (1) the CULTURA system and services and the two collections that have been used for testing and deploying the digital humanities research environment, and (2) an evaluation methodology and formative evaluation study with apprentice researchers. An evaluation model was developed which has served as a common ground for systematic evaluations of the CULTURA environment with user communities around the two test bed collections. The evaluation method has proven to be suitable for accommodating different evaluation strategies and allows meaningful consolidation of evaluation results. The evaluation outcomes indicate a positive perception of CULTURA. A range of useful suggestions for future improvement has been collected and fed back into the development of the next release of the research environment.

Keywords Virtual research environment · Digital humanities · Cultural heritage · Evaluation

1 Introduction

The interdisciplinary field of digital humanities is concerned with the intersection of computer science, knowledge management and a wide range of humanities disciplines.

Recent large-scale digitisation initiatives have made many important cultural heritage collections available online. This makes them accessible to the global research community and the interested public for the first time. In many cases, however, the full value of these heritage treasures is not being realised. Digital collections often lack features for deeper quantitative and qualitative analysis, and even very useful functions, such as the ability to annotate or bookmark content, are often not supported. After digitisation, these collections are typically monolithic, difficult to navigate, and can contain text which is of variable quality in terms of language, spelling, punctuation and consistency of terminology. Although there are digital content, tools, and services available, they are not necessarily useful or usable. This highlights the importance of continuous engagement and evaluation with users in order to ensure that design and development of digital humanities technologies correspond to the expectations and needs of their target audience, and to stimulate the use of these instruments. As stated in [1], "until analytical tools and services are more sophisticated, robust, transparent, and easy to use for the motivated humanities researcher,

C. M. Steiner (✉) · E.-C. Hillemann · A. Nussbaumer · D. Albert
Knowledge Technologies Institute, Graz University of Technology, Graz, Austria
e-mail: christina.steiner@tugraz.at

M. Agosti
Department of Information Engineering, University of Padua, Padua, Italy

M. S. Sweetnam
School of English, Trinity College, Dublin, Ireland

N. Orio · C. Ponchia
Department of Cultural Heritage, University of Padua, Padua, Italy

C. Hampson · G. Munnelly · O. Conlan
Knowledge and Data Engineering Group, Trinity College, Dublin, Ireland

Published online: 16 September 2014

it will be difficult to attract a broad base of interest…." As a result, digital collections often fail to attract and sustain broad user engagement, leading to limited communities of interest. Thus, important challenges still remain in the presentation of new digital humanities artefacts to the end user. Movement beyond self-contained, independent projects is needed, in order to create projects and flexible infrastructures usable for different collections.

Simple "one size fits all" web access is, in many cases, not appropriate in the digital humanities, due to the size and complexity of the artefact collections. Furthermore, different types of users have considerable differences in their knowledge of the collections, requiring varying levels of support, and every individual user has their own particular interests and priorities. Personalised and adaptive systems are thus important in helping users achieve optimum engagement with these new digital humanities assets. Improved quality of access to cultural collections, especially those collections which are not exhibited physically, is a key objective of the CULTURA project[1] [2,3]. Moreover, CULTURA supports a wide spectrum of users, ranging from members of the general public with specific interests, to users who may have a deep engagement with the cultural artefacts, such as professional and trainee researchers. To this end, CULTURA is delivering a corpus agnostic environment with a suite of services to provide the necessary support and features required for such a diverse range of users.

The CULTURA system has been tested and deployed with two contrasting digital humanities collections involving, respectively, textual material and images. This paper covers two main topics. (1) After providing an overview of existing virtual research environments (Sect. 2) the first integrated version of the CULTURA environment and the two test bed collections used are presented (Sect. 3). (2) Work done on evaluating the CULTURA environment is outlined, describing an evaluation model that has been developed based on related state of the art (Sect. 4) and how it has been applied in formative evaluation for user trials with apprentice researchers in the context of the two digital collections (Sect. 5). The evaluation outcomes are discussed in Sect. 6 and implications for further development of the CULTURA environment are drawn, also in terms of conclusions related to the applied evaluation method itself. Finally, overall conclusions and a panorama of future research are given (Sect. 7).

2 Virtual research environments

The tools and techniques of digital humanities allow massive cultural collections to be digitised, indexed, searched and combined with other digital libraries. Examples are the creation of digital archives of the transcribed 1641 Depositions testimony documents [4,5] and, respectively, of illuminated manuscripts with botanical illustrations included in the IPSA collection [6], as presented in more detail in Sect. 3 below. Large-scale programmes, such as Europeana[2], offer linked repositories that can facilitate search and discovery. The repositories tend to offer metadata about the artefacts, rather than a digitised artefact itself. This limits the form of research possible.

Recently, increasing effort has been made not only to make digital contents available, but also to create virtual research environments (VREs) that provide interpretative frameworks for making sense of cultural artefacts [7]. Such VREs support conceptualising, visualising and analysing information, as well as collaboratively working on it. They usually do not consist of one monolithic technology, but cover a collection of tools assembled in one place to assist research tasks and processes. Examples are Aus-e-Lit, a portal for the study of Australian literature [8], or the TextGrid environment for supporting researchers in the arts and humanities [9]. These environments incorporate tools for text search and analysis, archiving and reuse, collaboration and annotation. Commonly, they are developed for a particular digital collection and address a specific target audience, like professional researchers, research projects or institutions, or teaching communities. A major deficiency of most VREs is that they are designed with a 'one-size-fits-all' approach, thus making it difficult for users of varying experience levels to effectively make use of the content contained within.

In the field of historical, textual resources, most research environments are bespoke, designed to handle one particular type of text. This tends to limit the potential for reusing or repurposing tools. One notable exception to this is the Old Bailey Online[3] project. This important project makes available full transcriptions of the proceedings of the Old Bailey—London's main criminal court—from the period 1674–1913. This is a remarkably rich historical resource, and the project has ensured that this richness of content can be exploited by scholars, by designing an application programming interface (API) which allows the transcriptions to be interrogated in a rich variety of ways. This API has already been used, as part of the 'Data Mining with Criminal Intent' project[4] (a collaboration between the Universities of Sheffield and Hertfordshire, George Mason University, the University of Western Ontario, and the University of Alberta), to provide an integrated version of the Voyant suite of corpus analytic tools[5], and the Zotero bibliography management system[6].

[1] http://www.cultura-strep.eu/outcomes#2

[2] http://www.europeana.eu/.

[3] http://www.oldbaileyonline.org/.

[4] http://criminalintent.org/.

[5] http://voyant-tools.org/.

[6] https://www.zotero.org/.

This approach makes it possible to use a sophisticated and extensible range of tools on the material, and might be said to be the reverse of that adopted by CULTURA—it begins with the material, and adapts 'off-the-shelf' tools to fit. By contrast, the emphasis in CULTURA was the design of tools that could be generalised to a wide range of content types. In addition, the focus of the tools implemented so far by the Old Bailey Online project is primarily linguistic analysis. The corpus would clearly benefit from the sort of network analysis implemented by CULTURA, but at present this is not offered.

Network visualisation is a feature that is offered by the Electronic Enlightenment[7]. This resource gathers together almost 64,000 documents related to the Enlightenment. It is described as 'not simply an "electronic bookshelf" of isolated texts but a network of interconnected documents, allowing you to see the complex web of personal relationships in the early modern period and the making of the modern world.' It accomplishes this by allowing the user to look at either works or writers, and explore not just individual items, but the links behind them. This provides a very useful way of interacting with this material, but it is very much textually based, as compared with the visualisations provided by CULTURA. In addition, the project is based on large volumes of manually generated metadata, while CULTURA is able to enhance and enrich existing metadata.

In the History of Art field, Artstor[8] is a very well-known research environment, used both by professional researchers and students. The Artstor Digital Library is a nonprofit resource that provides over 1.6 million digital images in the arts, architecture, humanities and sciences. In Artstor users can search content by keywords or advanced search terms, view images and image data, zoom the images and create shared image folders, but arguably Artstor's most outstanding value is the breadth of its collection. Nevertheless, Artstor lacks those tools that make CULTURA so innovative, particularly the visualisation tool and the annotation tool (see Sect. 3) that allow a deeper engagement of the user with the collection. In fact, the visualisation tool helps users to quickly retrieve all the elements related to the images they are studying, while the annotation tool allows users to keep track of their research process, or simply to register their thoughts and impressions on a particular image.

CULTURA aims at building a novel type of VRE incorporating innovative information retrieval technologies and multidimensional adaptivity. The CULTURA system integrates a suite of intelligent services for guiding, assisting and empowering each individual user in their interactions with cultural artefacts. Thereby flexibility in terms of usage by a wide spectrum of users groups with their specific needs and in terms of reusability with different digital collections is provided, which characterises the innovative character of this research environment.

Because of its flexible approach to different types of corpora, and its ability to work with and enhance metadata of widely variable quality, CULTURA offers an interesting comparison with Europeana. Europeana contains a vast volume of material in a wide variety of formats. This material has been gathered from a range of sources, and has metadata of varying quality. At present, this complex range of material is exposed through a faceted search, which limits the scope for exploration and discovery within the collection. At present, the Europeana Cloud project[9] is seeking, amongst other things, to provide new tools, and to open up data to new types of tools. The CULTURA approach offers one example of how a collection like Europeana could be enriched for its users.

3 The CULTURA system

The CULTURA system consists of multiple distinct services all accessed via the CULTURA portal.[10] The services available are shown in Fig. 1 and include personalised search tools, faceted search tools, annotators, social network visualisation tools and recommenders. CULTURA seeks to offer a balance of personalisation and exploration tailored to each user's interests and experience. A specific service is triggered by a user's interaction with the CULTURA portal, with requests sent from the presentation layer to the service via its API. For example, when a person using the system is looking at one of the 1641 Depositions (see Sect. 3.1), entities from that document (people, places, etc.) are extracted from the data layer, and recommended Depositions based on these entities are calculated in the control layer by the recommender widgets. These recommendations are then rendered in the presentation layer for the user to view (see Fig. 2).

One of the key principles in the design of CULTURA is to ensure that users are in control of their experience. Users may interact with a user model tag cloud enabling them to scrutinise and adjust their user model. It is this model that underpins the personalisation offered by the recommenders, search and narratives. A user's individual model is constructed based on the interactions with the system. For example, if the user annotates an artefact using the annotation tool (FAST/CAT) [10], this indicates an interest in the entities contained within the specific parts of the artefact annotated.

CULTURA utilises Drupal[11] as the basis of its personalised portal, as Drupal provides numerous services that,

[7] http://www.e-enlightenment.com/.

[8] http://www.artstor.org/index.shtml.

[9] http://pro.europeana.eu/web/europeana-cloud.

[10] http://cultura-project.eu/.

[11] http://www.drupal.org/.

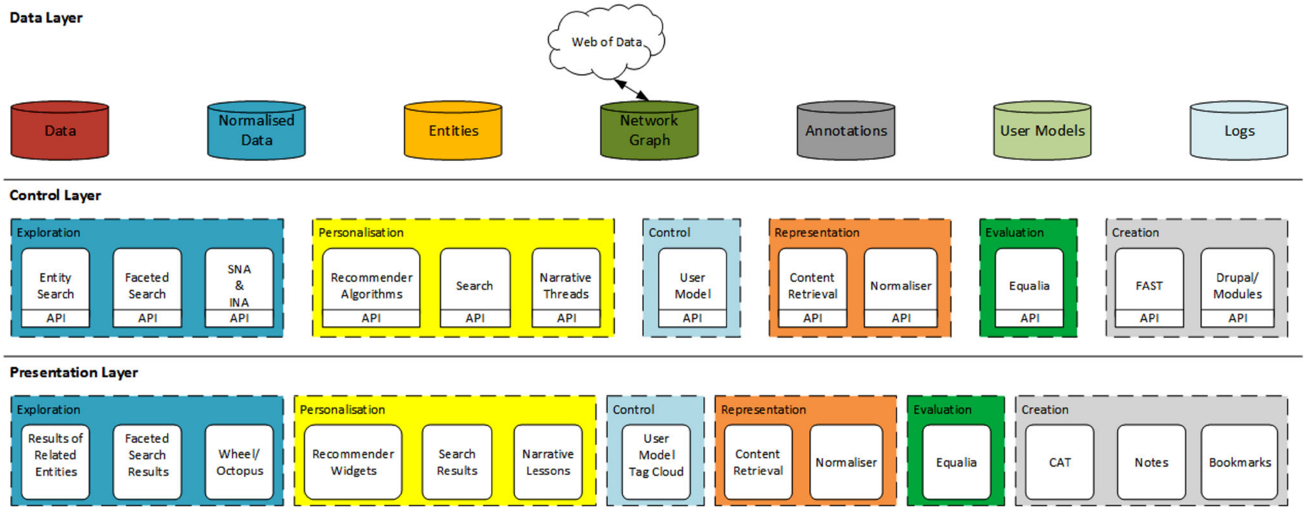
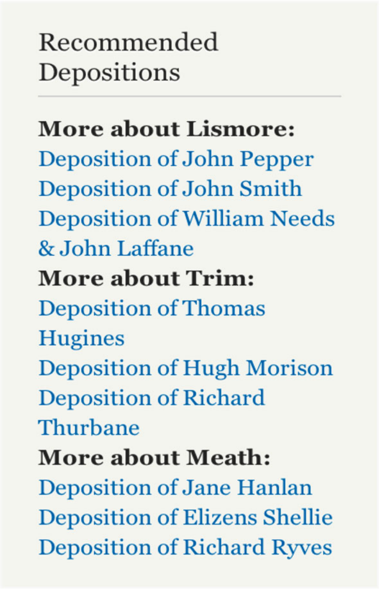


Fig. 1 The CULTURA architecture

while important to CULTURA, are not core research elements, such as user authentication and system-wide logging. Drupal also has an extensible architecture that allows new modules to be developed in order to extend or replace system functions. Hence, all services developed by CULTURA are implemented as Drupal modules, and when accessed by users, the responses from these services are displayed in an appropriate form, e.g. recommendations for related content as seen in Fig. 2.

The service-oriented architecture approach adopted by the CULTURA environment simplifies the integration process. The individual services offer powerful and novel functions in the areas of normalisation [11] and entity extraction [3]. These derived models of the content allow rich personalisation, via the recommenders and narratives to be supported.



Recommended Depositions

More about Lismore:
Deposition of John Pepper
Deposition of John Smith
Deposition of William Needs & John Laffane
More about Trim:
Deposition of Thomas Hugines
Deposition of Hugh Morison
Deposition of Richard Thurbane
More about Meath:
Deposition of Jane Hanlan
Deposition of Elizens Shellie
Deposition of Richard Ryves

Fig. 2 An example of the recommended content displayed to users within the CULTURA portal

Figure 1 highlights the various services that have been integrated in the control layer of the architecture. All that is required for each of these services is a well-defined API. In terms of the presentation layer, a user interacting with the entity-oriented search will be communicating with the entity-oriented search module in Drupal, which in turn accesses its bespoke API. Furthermore, because these tools support a parameterised launch, it also greatly simplifies the integration process. This is because all CULTURA services (visualisations, searches, etc.) can be rendered to users with the appropriate information, purely by passing the relevant identifiers to the service in the form of URL parameters. In this way, CULTURA can personalise both the selection and delivery of services to meet users' needs. Sections 3.1 and 3.2 describe the two cultural collections integrated into CULTURA, as well as the various services that the archives can access.

3.1 CULTURA and the 1641 Depositions collection

The *1641 Depositions* collection is held in the Library of Trinity College Dublin. It comprises more than 8,000 statements from witnesses and victims of the violence and atrocity that took place in the aftermath of the outbreak of the 1641 Rebellion in Ireland. The Depositions were recorded by government appointed commissions, and primarily record the experience of Protestant English settlers, the events they saw, or heard of, and the losses of property, possessions and money that they sustained. The Depositions are unparalleled in early modern Europe, and provide a unique window not only on the appalling events of the Rebellion, but also on the everyday and intimate lives of ordinary people and their efforts to make sense of the devastating disintegration of social order and neighbourly relations.

As part of a 3-year project, which commenced in 2007 and was funded by the Arts and Humanities Research Council (UK), the Irish Research Council for the Humanities and Social Sciences (Ireland), and the Library of Trinity College Dublin, the Deposition volumes, which were in a parlous condition, were conserved. High-quality digitisations of the Depositions were produced, and a team of three researchers transcribed the Depositions and captured extensive manually generated metadata, describing the occupation and address of the deponents, and the nature of the events recorded in each Deposition.[12]

From a technological perspective, the Depositions represent a textually rich digital humanities collection, which is characterised by noisy text, inconsistent sentence structure, grammar and spelling. The English language manuscripts contain rich metadata and descriptions of individuals, locations, events, social structures and contrasting/conflicting narratives. The digitised text of the Depositions and its associated metadata are stored in a MySQL[13] database that is accessed locally by the CULTURA Drupal environment. Because of the noisy text that is associated with the 1641 Depositions, a text normalisation process took place a priori [11]. The output of the normalisation process was added to the Depositions MySQL database and used to power normalised search over the Depositions, and to improve the entity relationship extraction performed using IBM's LanguageWare [12]. The process of entity relationship extraction created a graph of people, places, dates and events relating to the 1641 Depositions. Importantly, this entity graph was used in a number of key CULTURA services including social network analysis tools and visualisations, recommenders [13] and entity-oriented search [14]. Other important CULTURA services that operate over the 1641 Depositions include an annotation tool [10] which enables individuals and groups to create and share annotations.

3.2 CULTURA and the IPSA collection

The *Imaginum Patavinae Scientiae Archivum* (IPSA, Archive of images to support the study of scientific research at Padua University) collection is a digital archive of illuminated medieval and Renaissance codices, dating from the eleventh century. It contains astrological manuscripts and herbals with Latin, Paduan, and Italian language commentaries. In particular, herbals are manuscripts, which contain hand-drawn depictions of plants, such as trees, bushes or shrubs, and their parts, such as flowers or leaves, with a focus on their healing virtues. The IPSA collection contains mainly manuscripts written and illustrated by the Paduan School, and successive manuscripts produced in Europe under its influence. The online archive was created specifically for professional researchers in History of Illumination to allow them to compare illuminated images and to verify the development of a new realistic way of painting closely associated with the new scientific studies that were flourishing at the University of Padua in the fourteenth century, particularly thanks to the teaching of Pietro d'Abano [15].

Such manuscripts have the rare characteristic of containing high-quality and very realistic illustrations, because they were drawn from live specimens. The study of these manuscripts produced a number of scientific results on their content, which have been included in the collection. Thus IPSA is a combination of digitised images of the manuscripts and related metadata descriptions.[14]

The user requirements analysis with domain experts for the design and development of IPSA was conducted in 2002. A first complete prototype was made available to researchers in March 2003, a consolidated final version, revised using user comments, was released in July 2003 [6]. With the involvement in the CULTURA project, it was decided to open the archive to other categories of users, such as non-domain professional researchers, student communities and the general public. This new task required the identification of the needs, wishes and preferences of these new categories of users in order to define the required changes and improvements to IPSA.

Within CULTURA, IPSA metadata are shared in XML format, while high-resolution images of the illustration are loaded from an external server, due to copyright issues. The collection can be browsed using a keyword search or via a faceted browsing interface, both operating over the XML metadata. In addition, both the annotation tool and the social network visualisations (see Fig. 3) operate over the IPSA collection in the context of the CULTURA system. Due to the largely image-based character of IPSA, the normalisation component of CULTURA is not used with this collection.

3.3 The two collections in CULTURA

The aim of the CULTURA project is to pioneer the development of personalised information retrieval and presentation, contextual adaptivity and social analysis in a digital humanities context. This is motivated by the desire to provide a fundamental change in the way digital cultural heritage is experienced, analysed and contributed to by communities of interested individuals. These communities typically comprise a diverse mixture of professional researchers, apprentice researchers (e.g. students of history and art history), informed users (e.g. users belonging to relevant

[12] http://1641.tcd.ie/.

[13] http://www.mysql.com/.

[14] http://ipsa.dei.unipd.it/en_GB/home.

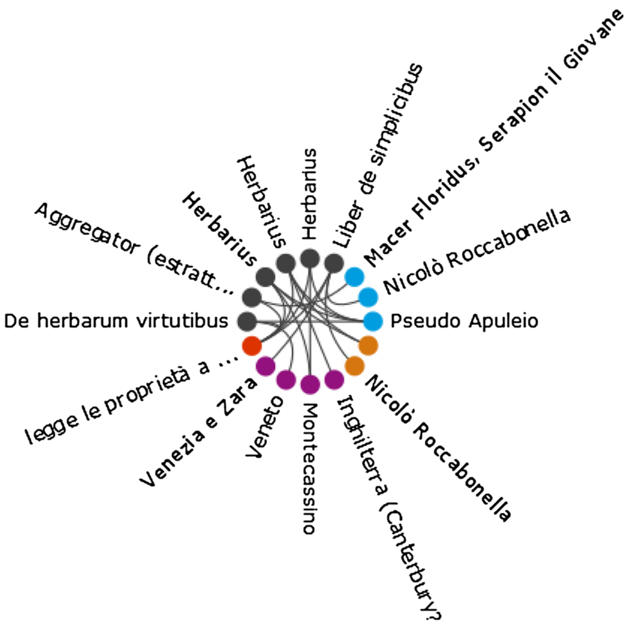


Fig. 3 Example visualisation based on social network analysis within the IPSA collection

societies, interest groups, or cultural authorities) and interested members of the general public.

From a technical perspective, IPSA and the 1641 Depositions represent two very different kinds of digital humanities collections. While the 1641 Depositions are basically textual documents, the IPSA collection is primarily image based, with substantial metadata available, which is also historically valuable as it captures the scientific processes which were prevalent during the creation of the original collection.

Notwithstanding these differences, CULTURA provides improved access to and powerful tools for exploring and researching both collections. The two collections share most of the CULTURA services. For instance, the annotation tool, which is used to annotate text within the 1641 Depositions, is used to annotate images and parts of images within the IPSA collection. These annotations can be used by an individual or made public to a specific group who may be working on the same topic. The social network analysis visualisations (see Fig. 3) used with the 1641 Depositions collection in CULTURA are also utilised with the IPSA collection, which highlights the generic nature of these Drupal modules and the effectiveness of a service-oriented architecture. Adaptation in terms of content recommendations is currently available only with the 1641 Depositions collection, but may easily be integrated for the IPSA collection on the basis of the available metadata.

The contrast in knowledge domain and structure of the IPSA and 1641 content collections demonstrate the broad applicability of the CULTURA methodology. Moreover, it highlights how the techniques delivered in CULTURA are not specific to an individual domain or collection, but can be of benefit to a wide range of digital humanities collections.

4 The CULTURA evaluation model

Hand in hand with the development of capable electronic information services and research environments for digital humanities, there is a need for comprehensive and scientifically sound evaluation of the quality of such information systems, in order to ensure that user needs are met and to inform further development. Current evaluation approaches can be categorised into three main types highlighting the targeted evaluation themes: user-oriented, system-oriented, and systematic [16,17]. A user-oriented evaluation approach pays attention to the user by examining users' requirements, behaviours and preferences, and their interaction, use and satisfaction with the digital library system or VRE in question. The main purposes of this type of evaluation are: verifying the quality of a product, detecting problems and supporting decisions [18]. System-oriented evaluation approaches focus on technological aspects and aim at investigating how well-advanced technology can be used for digital information representation and retrieval, measured for instance in terms of precision, recall and search time. This type of evaluation examines what happens in the informational environment external to the individual, while user-oriented evaluation examines the individuals' psychological and cognitive necessities and perceptions and how they affect information search and use. Systematic approaches address various levels or dimensions and thus may include user-oriented, as well as system-oriented evaluation goals. Different evaluation schemes and frameworks have been proposed in the literature, integrating a mix of dimensions and criteria from different disciplines (e.g. digital libraries, information retrieval, human-computer-interaction) and topics (e.g. content, engineering, user, environmental). Examples are the models suggested by Saravecic [19], Kovács and Micsik [20] or Zhang [21]. Fuhr et al. [22] established a framework for evaluation that integrates three existing evaluation models [19,20,23] under the umbrella of four categories (construct, context, criteria, and methodology) adapted from [19], which served for structuring and describing the evaluation process in a holistic manner.

One of these models is the so-called Interaction Triptych model [23,24], which analysed and described the interaction process with a digital library or research environment as a basis for deriving requirements and parameters for evaluation. Three main components of interaction are identified and captured by the model: the user, the content and the system. The analysis of the relationships between these components (i.e. user–system, user–content, and content–system), results in the following three evaluation aspects: usability (efficiency and effectiveness of user interaction with the system), usefulness (content usefulness and relevance to user tasks and needs) and performance (precision, recall, response time).

The novel technology of the CULTURA environment and the openness to a wide variety of content and users makes evaluation in CULTURA a challenging and multi-faceted task. On the one hand, CULTURA incorporates a range of different services, which required specific consideration in evaluation to get comprehensive outcomes on the quality of the system, and to identify aspects for further refinement of its individual components. On the other hand, the reusability of the methods and technology with different collections and the diversity of users taken into account necessitated an evaluation approach allowing researchers to select suitable methods for a specific evaluation task, while maintaining an appropriate level of comparability and generalisability of the evaluation results. This required a systematic approach of defining an evaluation methodology, which aimed not only at responding to the challenges in evaluating CULTURA, but also at offering potential for wider reuse in the evaluation of digital humanities research environments in general. An evaluation model was defined based on the existing state of the art and which could accommodate the service-oriented architecture and openness (to collections and to users) provided by a research environment like CULTURA [25]. The interactions in this kind of environment were analysed taking the Interaction Triptych model [23,24] as a starting point; i.e. the triple system, content and user have been conceptually identified as the basic components of the interaction process. The 'system' consists of the intelligent services as individual components, and of the system as a whole. The 'content' is given by the cultural heritage collections provided via the research environment—in our case the test bed collections 1641 Depositions and IPSA. With respect to 'users', four different user groups along the dimension of expertise are distinguished and addressed: professional researchers, apprentice researchers, informed users and members of the general public [26]. The evaluation aspects of the Triptych model were extended to address the quality axes specific to the research environment and its services, and to form a common ground for evaluation studies (see Fig. 4).

Usefulness of content refers, as in the original model, to the interaction between content and user: is the content relevant and suitable for the user? This relates to the question whether the digital collection supports the user's personal information needs and/or the information needs of the user group. A certain level of content usefulness is necessary for a meaningful evaluation of the other qualities.

Usability (also part of the original model) refers to the interaction axis between system and user: does the system allow users to effectively, efficiently, and satisfactorily accomplish their tasks? This relates to whether the communication and interaction between user and system are smooth and whether the system is easy to use and learn. It also includes aspects of the learnability, navigation and complexity of the system.

The system–user axis was complemented by *user acceptance*. Users may not necessarily have a positive attitude towards the system, even if it is technologically sound. User acceptance therefore addresses the specific question as to whether users consider the research environment and its services acceptable. Commonly, the following user acceptance aspects are distinguished [27]: perceived ease of use (which partly overlaps with usability aspects), perceived usefulness (this refers to the usefulness of the system and is to be distinguished from usefulness of content) and behavioural intention to use.

While these two axes of the interaction triangle are related to interaction evaluation, the aspect of *performance* (system–content axis) is usually not directly visible to the users and was not in the original scope of work on the Triptych model [24]. Performance is usually difficult to evaluate via user feedback and is commonly associated with information retrieval measures (precision, recall, response time). In CULTURA this evaluation axis was operationalised in terms of normalisation quality (quality and accuracy of the entity extraction and normalisation process) and network quality (accuracy of the data visualisations and the occurrence of probable inconsistencies between the entity data and the network visualisation). These evaluation aspects are not in the focus of the user-oriented evaluations presented in this paper. Rather, these aspects were evaluated in detail in system-oriented evaluations specifically addressing the related services, and are reported elsewhere (e.g. [28]).

In addition to the consideration of pairwise relations between the interaction components of the triptych, our evaluation model in particular also considers the ternary interrelation between all three components. In substance, this relation was addressed in terms of adaptation quality, visualisation quality and collaboration support (see Fig. 4) in the formative evaluation studies presented herein.

Adaptation quality refers to the interaction between system, content and user in terms of the system providing adaptive content recommendations tailored to the individual user: Is the adaptation provided by the CULTURA system appropriate and useful? This quality addresses users' perceptions of the helpfulness and benefit of system adaptation/recommendation received [29]. It can also be related to layered evaluation of adaptation [30], examining (a) whether user variables are correctly inferred and (b) whether adaptation decisions are appropriately taken.

Visualisation quality also addresses the interaction between all three model components: how do users feel about the visualisations of collection contents provided by the system? In the context of CULTURA this applies to the social network visualisations. Visualisation quality relates to user perceptions about the benefit of the visualisations provided and the user-friendliness of the visualisation tools.

Collaboration support is another quality at the centre of the evaluation model, relating to the collaboration between the users of a research environment and related to the content provided. It refers to the extent/quality to which users feel supported by the system to get in contact with each other and to share information on the collection content.

The evaluation model provides a sound theoretical basis for a systematic and comprehensive examination and validation of the novel functionalities integrated in CULTURA (e.g. visualisations), in addition to traditional evaluation topics on the overall system and of general interest (e.g. general usability assessment). It is suitable for application in formative, as well as summative evaluation. The model forms the mutual basis for setting up the design for all evaluation studies conducted in CULTURA over different collections and user groups. Although applying slightly different evaluation methods and strategies tailored to the user group, digital collection, and user trial setting in each case, a general level of comparability between evaluation outcomes over all user trials can be maintained. The common underlying conceptualisation of evaluation aspects thus makes the comparison and contrast of evaluation results from different user trials meaningful, even if they are gathered with different collections or from different user groups. While studies applying the same design and instruments offer the opportunity of quantitative and statistical comparison, with different evaluation designs investigating the same evaluation aspects nevertheless a comparison on a nominal or qualitative level is reasonable, in order to find out about aspects of the research environment that are generally positively or critically perceived and, respectively, to find out about specific issues that particularly apply only to a certain type of user group or to a certain digital collection.

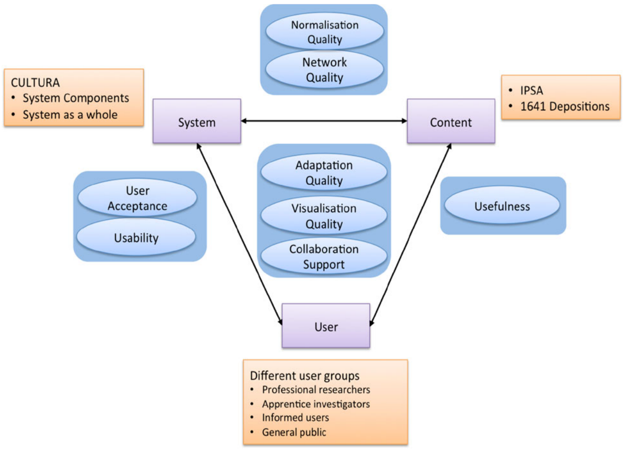


Fig. 4 The CULTURA evaluation model

In the remainder of this paper, a formative evaluation of the CULTURA system is outlined, which represents a concrete application of the evaluation model with a selected user group and for both test bed collections.

5 Evaluation of the CULTURA system

Since the CULTURA system is intended as a corpus agnostic environment suitable for different types of users, its evaluation needs to prove its benefits over different collections and user groups. In the following, empirical evaluations conducted on the two system entities are presented. CULTURA is designed to address the needs of a spectrum of users, ranging from the general public, who may be encountering the collections for the first time, to professional researchers who have worked extensively on the subject. Evaluation of the successive versions of the CULTURA implementation for the two collections took place over the three years of the project, and researchers worked closely with users from across the user spectrum. While evaluations with all user groups were conducted, here we present the evaluation methodology used and results gained with apprentice investigators as a selected user group in the formative evaluation phase. This group, together with professional researchers, is able to give the most in-depth and comprehensive feedback on the system with respect to the qualities of the evaluation model. Detailed feedback is especially valuable at a formative evaluation stage, when aiming at gathering evidence on the initial benefits and, even more important, information for further development. In the user trials presented below for both collections, the same general evaluation approach was taken.

5.1 General evaluation approach

A multi-method approach defined in line with the evaluation model was utilised with data collected from a variety of both quantitative and qualitative data sources. Data collection was carried out in three different ways: questionnaire, discussion and interaction logs.

5.1.1 Questionnaire

The survey instrument was developed in line with the evaluation model covering rating-scale items on all relevant evaluation qualities.

Usefulness of content was measured with two items on the relevance of the digital collection for individual and user group level.

For a general *usability* assessment, the System Usability Scale (SUS) [31] covering 10 items was used.

With respect to *user acceptance*, a scale (10 items in total) covering the main aspects of user acceptance according to the technology acceptance model [27] and already applied in the context of user acceptance research on digital libraries [32] was adopted.

The performance axis was not in the focus of this evaluation. Nevertheless, to capture user-centred aspects of *normalisation quality* in the context of the 1641 Depositions collection, two items asking for general level feedback on normalised search and entity-oriented search were used. Network quality was not explicitly addressed, since this aspect had been evaluated separately in a service-specific evaluation. Instead, user interaction with the provided visualisations (i.e. visualisation quality) was addressed in more detail (see below).

Adaptation quality was assessed by eight items on usage, usability aspects and perceived benefits of adaptive content recommendations provided by the CULTURA environment to users. While the usability assessment via SUS collected an overall assessment of usability for the whole system, the usability items in this subscale specifically addressed the adaptive recommenders. This complementary and more detailed consideration of usability-related aspects for the recommendation component was considered reasonable given the novelty of this kind of functionality in a research environment. Since recommenders were only available for the 1641 Depositions collection, the related questions were not presented in the case of the IPSA evaluation.

Visualisation quality was captured by nine questions on usage, aspects of the usability of the visualisation tools and their perceived benefits for users. Similar to adaptation quality, the collection of a specific assessment on usability aspects of the visualisations was included to ensure an in-depth consideration of the quality and perception of this novel type of exploration tool.

Collaboration support was measured with three items investigating the perceived support of collaboration with other users and the opinion about the annotation tool.

Five open questions were presented to collect qualitative feedback on the perception of the CULTURA system and features in a written form. The questionnaire was administered in separate online surveys for 1641 Depositions and IPSA. These surveys were made available to users through a personalised link within the CULTURA environment, and users completed them after they had spent time using the environment and its features.

5.1.2 Discussion

The moderated discussion with participants was designed in a semi-structured manner. Key questions in line with evaluation qualities were defined in order to add to questionnaire data and to gather more detailed user feedback on perceived benefits and potential issues for further improvement. To ensure that individual feedback was not influenced by the discussion, this discussion was done only after completion of the questionnaires.

5.1.3 Log data

User interactions with the CULTURA system were logged and examined. In the case of the 1641 Depositions user trial (which was carried out on a longer-term basis) for technical reasons, only data from the last month were available for further analysis. Log data analysis was done in accordance with the CULTURA evaluation model and its underlying evaluation qualities. This provided objective, quantitative data that complemented participants' self-reports, as available from questionnaire and discussion.

The evaluation of the CULTURA system was carried out in the context of university courses following a task-based approach. In these user trials, students (as apprentice investigators) were first introduced to the CULTURA environment and its functions. Subsequently, they were assigned research tasks and the CULTURA system was used to work on it. After task completion, students filled in the online survey and took part in the discussion.

This general evaluation approach was implemented in the form of two different but complementary evaluation strategies in the evaluation settings of the two collections.

5.2 Evaluation of 1641Depositions@CULTURA

5.2.1 Method

Participants The evaluation study in the context of the 1641 Depositions trial with Irish apprentice investigators involved in total 14 students, with only 11 (4 male, 7 female) of them completing the evaluation questionnaire. Participants were undergraduate, as well as masters students of History, Public History and Cultural Heritage, and Digital Humanities and Culture. The average age ($n = 11$) was 33.90 years (SD = 11.09), with individual ages ranging from 22 to 59 years. Students had advanced knowledge and experience of computers and computer applications in general.

Procedure Evaluation feedback was gathered from the apprentice investigators, who had spent 12 weeks working with the CULTURA system for the 1641 Depositions, and who had utilised it in the preparation of a number of different research exercises.

In using the 1641 Depositions collection, the following features were available to the users: content recommendations, social network visualisations, keyword search, search over normalised contents, faceted browsing interface, entity-oriented search, and the annotation tool.

Following the evaluation design described in Sect. 5.1, a mix of different methods and approaches was used, involving the gathering of both quantitative and qualitative feedback from users in accordance with the evaluation model and its evaluation qualities. Discussions for obtaining qualitative feedback were realised through a mix of focus groups, debriefing sessions, and one-on-one interviews after survey completion.

5.2.2 Results

Usefulness of content The usefulness of the 1641 Deposition collection content was assessed as very high, with an average score of $M = 6.09$ (SD = 0.58, Md = 6.0; see Fig. 5) on a scale with a possible score range from 1 to 7 (Note: due to the small sample size medians are reported in addition to arithmetic means). Overall, this excellent result can be interpreted as a high interest of students in the contents provided by the system. Users perceived the provided information as useful and relevant for their further studies and research.

Considering the log data from the last month of the user trial, participants visited on average 39 pages (SD = 45.51, Md = 18.0), of which $M = 26.42$ (SD = 39.71, Md = 5.0) were content pages presenting Depositions texts. Some students made quite extensive use of the system with more than 100 page visits.

Usability General usability scored moderately high with an average score of $M = 56.59$ (SD = 12.81, Md = 55.0) on a scale ranging from 0 to 100 with higher values indicating better results (see Fig. 5). Generally speaking, this moderate score on usability indicates an appropriate overall satisfaction in using the CULTURA system. Looking at individual items, the obtained results indicate that the learnability of the system is quite good, meaning that participants could easily accomplish tasks when working the first time with the CULTURA system. Most critically, the integration of the functions and tools was perceived as not very well achieved and not highly consistent. These results indicate that users did not find it easy to predict how the individual functions of the system work, which consequently leads to a slower progress operation as each one of the functionalities has to be learnt; this is a time- and attention-consuming process. Open feedback was, in general, very positive and identified issues concerned the technical implementation rather than the conceptual underpinnings of the tools. Almost all participants indicated that they would recommend CULTURA to a friend, mentioning usability aspects and research facilitations as the main reasons.

User acceptance User acceptance with its three main aspects—perceived usefulness, perceived ease of use and behavioural intention to use—was assessed quite positively with scores ranging from 4.91 to 5.43 (on a 1–7 scale, in each case). Results are depicted in Fig. 5. The best result was obtained for perceived usefulness ($M = 5.43$, SD = 1.24, Md = 5.5) indicating that participants perceived the CULTURA system as reasonably useful for supporting their research. This is in line with qualitative feedback, where nearly all students indicated seeing a potential benefit of using CULTURA for their research. Ease of use scored somewhat lower ($M = 4.91$, SD = 0.75, Md = 4.75), but still appropriately good. Participants had no problems in using the system and found it relatively easy to use. This user acceptance aspect was positively correlated with the overall usability score ($r = 0.66$, $p < 0.05$), as would be expected, since the latter also includes the aspect of learnability. For behavioural intention to use an average score of about 5 could also be identified ($M = 4.95$, SD = 1.69, Md = 5.50), indicating that participants envisage using the system in the future, if appropriate. Qualitative feedback confirmed the results on intention to use. Although some users stated that they were unlikely to continue to work on the Depositions, they still felt that the CULTURA approach has a high value and could be usefully extended to other corpora.

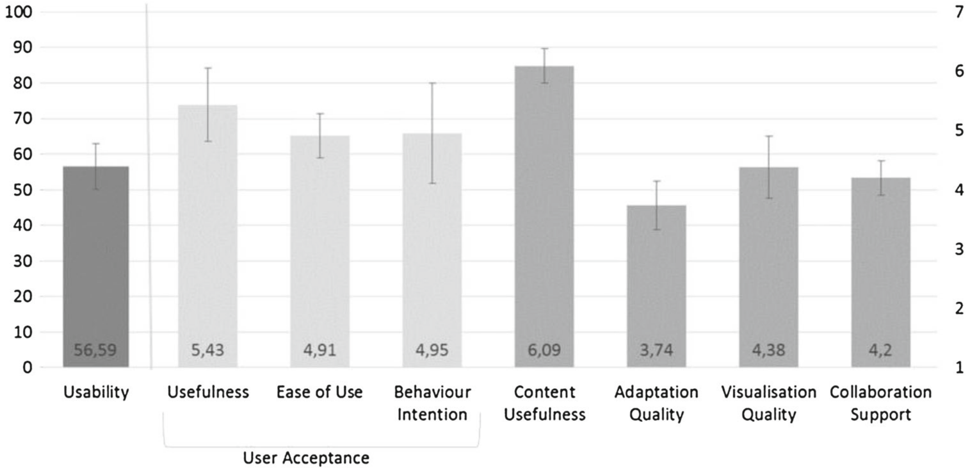


Fig. 5 Overview of evaluation results (mean scores and SD) on 1641Depsitions@CULTURA

Adaptation quality For adaptation quality an overall score was calculated; the item responses were also used to compute subscores on the estimated usage of content recommendations, their usability and their perceived benefit (possible score range 1-7). An average overall adaptation quality of $M = 3.74$ (SD = 0.82, Md = 3.71) was found, indicating a moderate to low result (see Fig. 5). Usability of the recommenders was assessed with medium quality, with an average score of $M = 3.93$ (SD = 0.89, Md = 4.0). The perceived benefit of adaptive recommendations was medium to rather low with $M = 3.68$ (SD = 0.62, Md = 3.75). The extent of their use was estimated by participants as medium to rather low ($M = 3.61$, SD = 2.01, Md = 4.0). This is completely in line with the log data. Examining these data, it became obvious that the recommendations provided by the system were rarely used by the students, with half of the participants having visited not one recommended Deposition and a mean number of $M = 2.14$ (SD = 5.52) content pages visited via the system's content recommendations. Overall, this quantitative feedback shows that students' experiences and perceptions of the recommenders are not as good as they might have been. Their rather low usage of the recommendations indicates that users did not really see the benefit they can receive from this tool. Nevertheless, qualitative feedback gathered in discussions confirmed the usefulness of recommendations as part of the research process, especially at the beginning when having low knowledge. This initial phase of system usage unfortunately could not be captured in terms of log data. However, users stressed the need for transparency by making explicit what was being recommended and why.

Visualisation quality. Similar to adaptation quality, for visualisation quality subscores (with possible score range 1–7) on estimated usage, usability and perceived benefit of the visualisation service were calculated from the questionnaire responses, and an overall score was derived. Overall visualisation quality scored $M = 4.38$ (SD = 1.05, Md = 4.0), which indicates a moderately good quality (see Fig. 5). The average estimated usage of the visualisation service was rather low with $M = 3.20$ (SD = 1.99, Md = 2.5). This result could also be confirmed by the log data showing that only 6 out of 14 students made use of the visualisations at all, most of them only a few times. The result for usability is satisfactorily good, with half of the participants scoring with 4.5 or higher (i.e. *Median*) and on average with $M = 4.58$ (SD = 0.71). The best result could be found for the perceived benefit with a mean score of $M = 5.38$ (SD = 0.79, Md = 5.25). Summarising, students generally agreed that visualisations provide relevant information, are moderately easy to use, and not too complex. Additionally, they perceived the visualisations provided by the system as being able to support them in better understanding the resources provided by the collection. Discussions confirmed these results; users were on the one hand intrigued by the visualisations and expressed a keen sense of the potential usefulness. However, on the other hand, they pointed to the need for more flexible visualisations in terms of being able to move between several texts and visualisations at the same time or of visualising more than one Deposition at a time.

Collaboration support For collaboration support a medium quality could be identified ($M = 4.20$, SD = 0.63, Md = 4.0; see Fig. 5). This result indicates that participants had some idea but were not totally convinced whether the system can support users to communicate with each other and collaborate on research tasks. The explicit assessment of the annotation tool and its usefulness (as a tool and indicator for collaboration) indicated a good quality with $M = 5.20$ (SD = 1.03, Md = 5.0). This means that students are aware that the opportunity to create and share annotations is useful. However, log data showed that there were only three persons who had annotated content pages. Thereby, the number of annotations taken by individual participants ranged from 1 to 6. Nevertheless, when explicitly asked to identify the single most useful feature of the CULTURA environment in the open feedback section of the questionnaire, students highlighted the annotation feature, as well as other features (like the search over normalised contents, the faceted search, and the visualisations).

Normalisation quality The ratings on normalised search and entity-oriented search were positive, resulting in an average score of $M = 5.4$ (SD = 1.17, Md = 5.5) for entity-oriented search and $M = 6.40$ (SD = 0.97, Md = 7.0) for search over normalised contents. These quantitative results indicate that users generally found both functionalities useful for their research work. On average, students made about 10 text searches ($M = 9.79$, SD = 12.67), half of them ($M = 5.14$, SD = 8.95) were carried out over normalised contents. Open feedback from discussions showed that in practice, users were very pleased with the ability to search over normalised data. Concerning the entity-oriented search interface, though users saw high potential and value in this feature, they pointed to two problems: the accuracy of automatically generated metadata and the rather confusing and slow interface.

5.3 Evaluation of IPSA@CULTURA

5.3.1 Method

Participants In total 110 Italian apprentice investigators took part in this study and completed the evaluation questionnaire. This sample included undergraduate, as well as master students attending university courses of History and

Preservation of Cultural Heritage and of Management of Archival and Bibliographic Heritage. Gender distribution was 81 (74 %) female and 29 (26 %) male students. The average age of participants was 21.57 years ($SD = 2.38$) with a range of 18–29 years. Students had average computer literacy and experience.

Procedure Contrary to the evaluation of 1641Depositions@CULTURA, where the user trial not only involved a small group of apprentice investigators but also consisted in a longer period of contact and interaction with the CULTURA system, the evaluation of IPSA@CULTURA was characterised by a shorter-term interaction and engagement with the system by a large sample of students. Since a large sample makes very intensive interaction and exchange with individual participants difficult and also due to a limited number of computer workstations available, the user trial was carried out by dividing the students into smaller groups of about 20-30 students.

In using the IPSA collection, the system features included social network visualisations, faceted browsing interface and the annotation tool. Data collection followed the general procedure described in Section 5.1 using a mixed-method design including questionnaires, focus group discussion and log data.

5.3.2 Results

Usefulness of content Content usefulness was assessed at a medium level with $M = 4.26$ (SD = 1.46), as depicted in Fig. 6. This is mostly due to the fact that the IPSA collection served rather as a showcase on preservation and archival management of cultural heritage; the young researchers were not specialised in the domain of the collection, but had a more general focus on cultural heritage and digital humanities. This is in line with results obtained from the discussions, where students pointed to their limited knowledge of the collection, but expressed their interest in the use of CULTURA with other collections.

Users visited on average 20 pages ($M = 20.44$, SD = 12.41), of which 7 ($M = 7.40$, SD = 6.76) were content pages presenting illuminations.

Usability From the responses collected in the evaluation of IPSA@CULTURA a usability score of $M = 65.39$ (SD = 14.17) was determined (see Fig. 6), indicating satisfactorily good usability. Looking at individual items, participants assessed the consistency of integrated functionalities within the system as quite good. Furthermore, they found the learnability of the system appropriate. A lower score was found for the item on potential future use; it seems that students did not consider using the system for their research or studies as highly relevant (which may also be related to the results on usefulness of content), or at least they are unable to recognise the potential support of using the system. This is also reflected in the qualitative feedback collected, where some students pointed out that a tutorial or training material explaining the most important CULTURA functions in more detail would be useful to better understand the potential relevance and usefulness of using the CULTURA system to support their research.

User acceptance Ratings on perceived usefulness, perceived ease of use, and behaviour intention to use had been collected as aspects of user acceptance. An overview of the mean scores for these subscales is given in Fig. 6. The best result was obtained for ease of use with $M = 4.95$ (SD = 1.26), which is also closely linked with usability, resulting in a positive correlation of $r = 0.67 (p < 0.01)$ between the two scores. Students who judged usability as high also

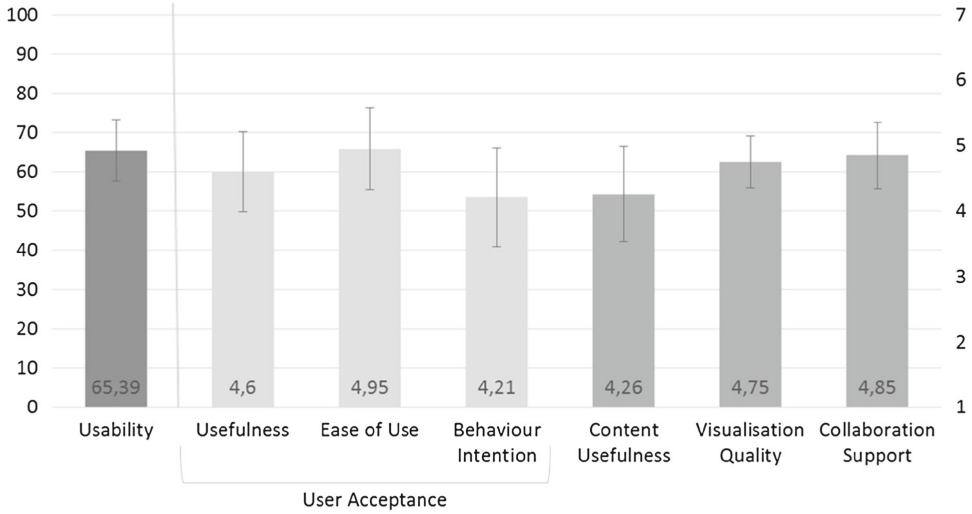


Fig. 6 Overview of evaluation results (mean scores and SD) for IPSA@CULTURA

assessed the ease of use in the user acceptance scale as being high. This result confirms that students perceived the system's learnability as quite good. Perceived usefulness was rated with an average score of $M = 4.60$ (SD = 1.13), thus indicating a medium to good result. This is confirmed by qualitative feedback, where students highlighted that CULTURA is a useful tool for their research, especially with regard to cultural heritage artefacts and their availability. Regarding behaviour intention to use the system, a mean score of $M = 4.21$ (SD = 1.41) was found, indicating that students had only a moderately high intention of using the system. This may be due to the fact that students, as expressed in open comments, were not specialised and had limited knowledge of the content and therefore did not see great benefit in using the system with this specific collection in the future.

Visualisation quality Overall visualisation quality scored with a mean of $M = 4.75$ (SD = 0.80) indicating a medium to good result (see Fig. 6). The extent of estimated usage of the visualisations provided by the CULTURA environment was modest with a mean score of $M = 3.57$ (SD = 1.33). This corresponds to the log data from the user trial, which shows that visualisations for illuminations were accessed rather rarely, about 4 times on average ($M = 4.07$, SD = 3.53). With respect to usability of the visualisation tool, a mean score of $M = 5.05$ (SD = 1.00) was identified, arguing for satisfactorily good usability. The perceived benefit of the visualisations was also quite good and reached a mean score of $M = 5.09$ (SD = 1.02). Overall, participants acknowledged the benefits the visualisations can bring, to gain new insight on the digital contents, as well as their usability, while their use of visualisations was rather low. Open responses on the visualisations indicate that this feature was appreciated by students and was perceived as useful. However, users wished for a more detailed explanation in order to get a better understanding of the graphical illustration and how to interpret it.

Collaboration support The mean score for collaboration support with $M = 4.85$ (SD = 1.01) was appropriately good (see Fig. 6). Participants considered the CULTURA system as capable of supporting collaboration among users to a satisfactory extent. The annotation tool was assessed as being of good quality with $M = 5.16$ (SD = 1.20). On average students created more than 5 annotations ($M = 5.30$, SD = 2.55). In addition, more than one annotation was commonly made within a single IPSA illumination. Open responses on collaboration support and the annotation tool show that these features were perceived very positively. However, further refinements in terms of identifying authors and a rating system for annotations were suggested, which would allow improvements to both the interpretation and validity of annotations.

6 Discussion of evaluation outcomes

6.1 Implications for the CULTURA environment

The two user studies presented in this paper are characterised by different evaluation settings and strategies addressing different subgroups of the user group of apprentice researchers in the context of two contrasting digital collections. Nevertheless, the evaluation model gave ground to a common evaluation approach and established overall comparability between the user trials. By consolidating the general results of the two evaluations, aspects of the CULTURA system that are generally positively perceived or looked at critically by the involved user cohorts can be identified.

Content usefulness of the digital collection provided via the CULTURA environment was perceived quite differently in the two user trials: while apprentice investigators assessed the 1641 Depositions collection contents as highly relevant and interesting to them, students in the user trial on the IPSA collection perceived the collection content only as moderately relevant. Both cohorts acknowledged the potential reuse of CULTURA with other corpora as highly beneficial.

For both cohorts overall usability of the CULTURA environment turned out to be satisfactory, with students in the IPSA trial assessing the system more positively. Considering the most critically assessed individual usability items, a somewhat different picture emerged for the two studies: In the IPSA trial users seemed to have particular difficulty in considering a frequent future use of CULTURA, which also mirrors the moderately perceived relevance of the collection in this trial. In the 1641 Depositions trial, the integration and consistency of the system were especially highlighted as a critical issue needing further improvement. In this regard, it has to be taken into account that the system provided a broader range of different services for the 1641 Depositions collection than those available for the IPSA collection. This is largely due to IPSA being an image collection rather than a textual collection like the 1641 Depositions.

A comparison of results for user acceptance aspects, which in principle all scored moderate to good, confirms the previous outcomes: in the IPSA trial the system was evaluated as easier to use, but behaviour intention to use the system was rather low; in the 1641 Depositions trial, in contrast, the perceived usefulness and behavioural intention to use were slightly more pronounced. Additional explanation of the CULTURA services in the system and demonstration (through engagement with users) of how they can be used in research and exploration is to be assumed to further improve perceived usefulness, as well as usability.

With respect to the evaluation of the visualisation tools, a moderate to good quality could be identified in both evaluations. Interestingly, in both groups apprentice investigators did not extensively use the visualisations, but nevertheless evaluated their usability and, especially, the potential benefits users can derive from the visualisations as good. In the context of the 1641 Depositions collection it became clear that students would be interested in a further extension of the visualisations in order to increase their added value for gaining insight into the research process.

The annotation tool was assessed consistently positive. Collaboration support was perceived as medium to good by apprentice investigators with the IPSA collection, and therefore, as better than for the 1641 Deposition collection, which had only a medium result. In total, these results indicate that there is potential to further extend the system with additional functionality supporting collaboration, e.g. with chat or discussion forum features.

The possibility of searching over normalised contents of the 1641 Depositions collection was highly appreciated, which underlines the importance of normalisation and argues for normalisation quality from a user-centred perspective.

Overall, qualitative feedback was quite positive and students found the CULTURA environment interesting. In both trials, students expressed their interest in having CULTURA available for other types of collections. Suggestions for additional features made by students in the IPSA context included, among others, integration with social networks, but also different language versions of the environment and additional help/tutorials; students' suggestions in the 1641 Depositions trial addressed mainly the possibility of organising and exporting their data, like bookmarks, annotations and searches.

The recorded and analysed log data for both studies reflects well the two different evaluation approaches taken in the two trials. Although for 1641Depositions@CULTURA log data were available only for a limited time span of the overall duration of the user trial, the analysed data set clearly shows that this user trial implemented a more intensive interaction with the system (in terms of page visits, content pages accessed and searches conducted), with a smaller group of users over a longer period of time. This was in contrast to a shorter and less intensive interaction with a large number of participants for IPSA@CULTURA. Despite this, individual features like the visualisations or bookmarking were used similarly scarcely in both trials. The annotation functionality was used more intensely in the IPSA trial, which may be explained by the fact that the tasks in this study explicitly requested the creation of annotations. It needs to be taken into account, though, that for the Depositions collection log data only reflects an extract of the whole usage period; the usage figures obtained thus need to be considered with caution, since some students might have concentrated their usage predominantly within the first two months for which no interaction data were available.

Comparing the obtained and consolidated evaluation results and implications with broader evaluation outcomes from user trials with other user groups and even with other digital collections, useful conclusions can be drawn for informing further development. Concretely, the implications drawn from the user trials with apprentice researchers have been checked against and compared with outcomes of other evaluations conducted with professional researchers (e.g. [25]), informed users and members of the general public (e.g. [33]). Overall, in all formative user studies a generally positive assessment of the CULTURA research environment could be obtained indicating high user interest, satisfactory usability and user acceptance and a positive perception of the CULTURA services, with questionnaire scores and qualitative feedback nevertheless indicating some room for further improvements. Agreement on issues reported and improvements suggested could be identified across user groups. For example, users from all categories asked for an additional explanation of the system features in the help section of the environment. There was consensus among all user groups about the usefulness of the annotation tool and the need to further improve this functionality. With regard to the visualisations, feedback from all studies confirms the innovative and appealing character of this kind of service. User groups more experienced in digital humanities consistently mentioned the visualisation of several content items at a time as a desirable feature. User groups with lower expertise with the digital collection and VREs (i.e. informed users and members of the general public) mentioned that it would be nice to have the possibility to provide feedback on the system while using it. In the context of the 1641 Depositions they furthermore indicated that they would also like to access the normalised text, since they had difficulties in understanding the original Deposition texts.

The aggregation of the main results obtained from the totality of formative evaluation studies informed further development, thus closing the feedback loop from evaluation back to implementation work. As a result, the subsequent release of the CULTURA system and its components incorporates a whole range of improved or additional functionality based on end user feedback: the visualisation of multiple artefacts, extended annotation functionality, the inclusion of live text normalisation, the possibility to save searches, improved help functionality, as well as the integration of a tool for providing on-line evaluation feedback.

6.2 Implications for the evaluation methodology

The evaluations in the context of the two collections were based on the same common evaluation approach and model, but were obviously quite different in nature. In the sequel, the lessons learned from the evaluation strategies are discussed.

In both cases, students as apprentice investigators were involved in the evaluation studies, with the user trials being integrated in the students' curriculum. In the case of the 1641 Depositions collection, the user trial consisted of a longer term involvement of a small group of users and intensive engagement with the CULTURA system. In the context of the IPSA collection, the user trial consisted of a shorter contact with a large number of participants. In addition, users differed in their age, with participants from the first being considerably older than those from the second evaluation. This difference was also reflected in user expectations on using the digital collections and CULTURA in the future: students in the first year of their studies, like in the IPSA trial, commonly do not yet have a clear idea of the tools they will need for their career.

Another important difference between the two samples was that English was the participants' native language only in the evaluation of 1641Depositions@CULTURA, while participants in the user trial on IPSA@CULTURA spoke Italian. This might have had an influence on user interaction with and assessment of the CULTURA environment, which was only available in English.

An evaluation strategy as used with 1641 Depositions provides the advantage that users acquire sufficient knowledge of the system and, after intensive interaction, are able to provide in-depth feedback on their perception of and experience with the research environment. Such intensive engagement with users is only possible with small numbers of users, which in turn complicates valid quantitative data analysis. This kind of approach involves a high level of workload and time effort from evaluators and participants; therefore, it sometimes might not be feasible to apply this kind of strategy in evaluation practice. An evaluation strategy as used in the context of the IPSA collection, in contrast, is easier to organise and deploy. A more short-term interaction with the system, though, might be considered more as a snapshot of user experience and feedback and may lead to user assessments being somewhat confounded with learnability, in the case users have not had the chance to gather sufficient experience in handling the system. If an appropriate exposure to the system is possible even in the context of a short-term user trial, participants will be able to provide valuable feedback, but probably not on the level of detail as in case of the first evaluation strategy.

Overall, the differences in the evaluation strategies applied make them complementary in terms of a comprehensive and mixed-method evaluation of CULTURA in general. By bringing together the outcomes from both studies and strategies, more conclusive evidence could be drawn on the overall quality of the system and services. In addition, aspects for future improvement of the research environment could be identified, with the potential of supporting users regardless of the digital collection.

The evaluation methodology and model have proven a suitable ground for establishing a common evaluation approach for both user trials and the involved evaluation strategies. It has to be noted that, in principle, different evaluation instruments are more or less suitable for specific evaluation settings. Interviews or focus groups, for example, are feasible only with smaller samples due to the high workload involved, while surveys are especially suitable also for larger sample sizes. From small samples, it is possible to gather in-depth qualitative feedback, but quantitative analysis of survey responses is somewhat problematic. With large samples, the detail in feedback collected will necessarily be somewhat shallower, but questionnaire scores obtained are suitable for quantitative and statistical analysis. When translating the evaluation model into a concrete study design for application in different user trials, a matching between data collection method and evaluation setting, therefore, always needs to be taken into account. Likewise, the extent and granularity level of feedback that different end user groups are able to provide on a research environment and on specific features also need to be accounted for when operationalising the evaluation model in terms of concrete assessment instruments. The mix of methods applied in our case was appropriate for data collection in both evaluation settings, since it incorporated instruments accommodating the specific conditions given in both user trials and meaningfully complementing each other.

The consideration of different evaluation aspects along the axes of interaction components, as captured by the evaluation model, resulted in a highly suitable approach for gathering comprehensive and conclusive evidence on the quality of the system, about specific benefits and, even more importantly, about drawbacks to be fed back to development. The evaluation model also provides the freedom to select and address only the evaluation aspects that are of relevance given a certain evaluation setting (e.g. disregarding adaptation quality for the IPSA collection), while nevertheless maintaining general comparability between user trials with respect to qualities addressed in those trials.

7 Conclusion and outlook to future research

This paper introduces the CULTURA research environment, which provides innovative functions to support research and exploration of digital heritage collections. Two evaluation studies are presented that were conducted in the context of the two test bed collections using the CULTURA system. The evaluation approach taken accommodates the different and complementary evaluation strategies applied, which were aligned to a common underlying evaluation model, and ensures a general comparability of results.

The user trials presented involved in total 121 apprentice investigators. Although this sample is not necessarily a definite and representative sample of the community of humanities apprentice researchers, in general, it represents a reasonably large set of opinions and assessments in the context of the test bed collections. The user group of apprentice investigators together with professional researchers is considered important target audiences, who are able to provide evaluation feedback with a high level of detail. User-centred evaluations involving the user groups actually addressed by a research environment are key, because only those people who are in a position to benefit directly from such a research environment or information service are able to take this field of research and development forward [1]. Since the CULTURA system is not intended to be an environment for researchers only, but targets the user spectrum along the whole dimension of expertise, evaluation studies other than the two presented in this paper also involved other user groups [26]. One study, for example, addressed secondary school students as members of the general public [33]. This is especially interesting, since there is increasingly high potential seen in using VREs in digital humanities as teaching instruments [6], which is relevant not only for academic, but also for elementary and high school education.

The evaluation outcomes on the different qualities of the evaluation model provided targeted information on aspects and potential for further refinement or extension of specific features of the CULTURA environment. The results obtained from all user studies on the same system release were consolidated to derive implications for further development. A range of changes was implemented in a new system release in the meantime. These were evaluated positively in more recent user trials and were singled out as being especially valuable in terms of building user comfort and confidence in the CULTURA environment. The evaluation results of the formative evaluation studies presented in this paper also provide benchmark data for comparison with summative evaluation outcomes with the aim of demonstrating progress made and the overall benefits and quality of CULTURA as a novel research environment.

In summary, the CULTURA system includes a set of components that provide powerful and innovative supporting tools for different user interactions with cultural heritage contents. Its flexibility and reusability with diverse kinds of digital collections make CULTURA a research environment that has the potential of opening up and facilitating unified access to a whole range of different cultural collections and for the complete spectrum of end users. Meanwhile, a new content collection comprising witness statements from the 1916 Rising[15] has been incorporated in CULTURA. The inclusion of this material demonstrates the generalisability of the CULTURA environment, and its component tools, and it is also an important element of the sustainability plan for the environment.

The CULTURA system was designed from the outset to meet the requirement of a corpus agnostic research environment, i.e. to be independent of any specific collection and rather work alongside them. This means that, although CULTURA was evaluated within the context of the two digital collections, it is not designed specifically with either of these in mind, but rather as a flexible and supporting framework capable of integration with a wide range of digital collections. The environment, developed with a particular set of services, may not have all those capabilities implemented ideally or used at all within an application on a specific given digital collection. This may be due to the nature and domain of the digital collection, as well as to technical, pedagogical, or pragmatic reasons. This has to be taken into account when considering individual evaluation studies in a very concrete application setting, given a certain collection managed by the CULTURA system, and a selected user group.

CULTURA is intended as a corpus agnostic and user group-independent research environment. Consolidating the results over different evaluation studies allows researchers to discover issues and implications that are of general interest, as well as aspects regarding the overall quality of the environment. Taken together, different evaluation studies on CULTURA will prove its general usefulness and significance for empowering and guiding users in their interaction with digital humanities collections. The evaluation methodology thereby constitutes a common reference point for specifying the data collection instruments and for comparing and generalising results. The evaluation model gives ground to a comprehensive and in-depth gathering on the relevant evaluation aspects of the system, and provides freedom in terms of the concrete instantiation through an evaluation design and the evaluation strategy applied, as well as in terms of selecting the qualities relevant for a certain evaluation setting. It is even open for the inclusion of additional, complementary evaluation qualities that may be associated with the interaction axes, as appropriate in light of future development of CULTURA or, respectively, a different VRE. In fact, with the maturation of the CULTURA technologies in the final development phase, a further refinement of the evaluation model for summative evaluation evolved in order to appropriately capture and assess all the different functionalities of the system and to enable systematic information gathering on their quality and benefit for users. The evaluation qualities defined in the CULTURA evaluation model were revisited in the light of the latest technical developments, and the qualities were refined or elaborated in more detail, where appropriate.

The evaluation model is considered to have high potential for reuse in other research environments. Aside from aspects of general usability, acceptance and usefulness of contents, in particular tools for visualising the contents of a digital

[15] http://cultura-project.eu/1916/.

collection to support understanding and exploration are relevant also with other systems. Besides, features enabling a personalisation of user experiences to individual needs and expertise and supporting collaboration and exchange between groups of users are becoming increasingly important in electronic information services for cultural heritage in general. This makes the evaluation qualities specified in the CULTURA evaluation model applicable on a more general level. Naturally, different aspects will be focused on different degrees and addressed by diverse technical approaches in individual research environments. The evaluation model provides a valuable starting point for identifying the axes and topics that are of interest in a new evaluation project and for defining the actual evaluation design and instruments to be applied. A broader reuse of the evaluation model in digital humanities would also facilitate comparison between evaluation results across different research environments.

The evaluation model has also been used as a basis for the development of an evaluation service [34], aimed at supporting evaluators in planning, carrying out and analysing evaluations. Through explicitly specifying the quality model underlying an evaluation within the service, data collection can be systematised, automated reports can be derived, and data gathered via different collection modes can be triangulated, with the evaluation model as a reference base.

Acknowledgments The work reported has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement no. 269973, and would not have been possible without the close collaboration between all CULTURA partners. http://www.cultura-strep.eu/partners.

References

1. Borgman, C.L.: The digital future is now: A call to action for the humanities. Digital Human. Quart. **3**(4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html (2009). Accessed 3 Oct 2013
2. Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., Wade, V.: The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In: Proceedings of the 4th International Euromed Conference, pp. 668–675. Springer, Heidelberg (2012)
3. Hampson, C., Lawless, S., Bailey, E., Yogev, S., Zwerdling, N., Carmel, D.: CULTURA: A metadata-rich environment to support the enhanced interrogation of cultural collections. In: Dodero, J.M., Palomo-Duarte, M., Karmpiperis, P. (eds.) Proceedings of the 6th Metadata and Semantics Research Conference, pp. 227–238. Springer, Berlin (2012)
4. Clarke, A.: The 1641 Depositions. In: Fox, P. (ed.) Treasures of the Library, pp. 111–122. Trinity College, Dublin (1986)
5. O'Regan, D., Sweetnam, M., Fennell, B., Lawless, S.: A collaborative linguistic research interface for the 1641 Depositions. Poster presented at Digital Humanities 2011. Stanford, CA (2011)
6. Agosti, M., Benfante, L., Orio, N.: IPSA: A digital archive of herbals to support scientific research. In: Sembok T.M.T. (ed.) Proceedings of the International Conference on Asian Digital Libraries ICADL 2003. LNCS vol. 2911, pp. 253–264. Springer, Berlin (2003)
7. Bellamy, C.: The sound of many hands clapping: Teaching the digital humanities through virtual research environments (VREs). Digital Human. Quart. **6**(2). http://www.digitalhumanities.org/dhq/vol/6/2/000119/000119.html (2012). Accessed 3 Oct 2013
8. Hunter, J., Gerber, A.: The Aus-e-Lit project: Advanced eResearch services for scholars of Australian literature. VALA 2010. Melbourne, Australia. http://www.itee.uq.edu.au/eresearch/papers/2010/VALA2010_Hunter.pdf (2010). Accessed 3 Oct 2013
9. Neuroth, H., Lohmeier, F., Smith, K.M.: TextGrid—virtual research environment for the humanities. Int. J. Digit. Curation **6**(2), 222–231 (2011)
10. Agosti, M., Conlan, O., Ferro, N., Hampson, C., Munnelly, G.: Interacting with digital cultural heritage collections via annotations: the CULTURA approach. In: Proceedings of the 2013 ACM Symposium on Document Engineering, pp. 13–22. ACM, New York (2013)
11. Lawless, S., Hampson, C., Mitankin, P., Gerdjikov, S.: Normalisation in historical text collections. In: Proceedings of Digital Humanities 2013, pp. 507–509. Lincoln, Nebraska (2013)
12. Yogev, S., Roitman, H., Carmel, D. Zwerdling, N.: Towards expressive exploratory search over entity-relationship data. In: Proceedings of the 21^{st} International Conference Companion on World Wide Web, pp. 83–92. New York (2012)
13. Hampson, C., Bailey, E., Munnelly, G., Lawless, S., Conlan, O.: Dynamic personalisation for digital cultural heritage collections. In: Proceedings of the 6th International Workshop on Personalized Access to Cultural Heritage. Rome (2013)
14. Carmel, D., Zwerdling, N., Yogev, S.: Entity oriented search and exploration for cultural heritage collections: the EU CULTURA project. In: Proceedings of the 21^{st} International Conference Companion on World Wide Web, pp. 227–230. New York (2012)
15. Mariani Canova, G.: Per Cultura: le immagini dei manoscritti della scienza a Padova dal Medioevo al Rinascimento. In: Atti e Memorie dell'Accademia Galileiana di Scienze, Lettere ed Arti. vol. CXXIV, pp. 81–90 (2011–2012)
16. Saracevic, T.: Digital library evaluation: toward an evolution of concepts. Libr. Trends **49**(2), 350–369 (2000)
17. Zhang, Y.: Developing a holistic model for digital library evaluation. Dissertation. The State University of New Jersey. (2007)
18. De Jong, M., Schellens, P.J.: Reader-focused text evaluation: an overview of goals and methods. J. Bus. Tech. Commun. **11**(4), 401–432 (1997)
19. Saracevic, T.: Evaluation of digital libraries: An overview. In: Agosti M., Fuhr N. (eds.) DELOS Workshop on the Evaluation of Digital Libraries. Padua, Italy. http://www.scils.rutgers.edu/~tefko/DL_evaluation_Delos.pdf (2004). Accessed 7 Oct 2013
20. Kovács, L., Micsik, A.: The evaluation computer: A model for structuring evaluation activities. In: Agosti, M., Fuhr, N. (eds.) DELOS Workshop on the Evaluation of Digital Libraries. Padua (2004)
21. Zhang, Y.: Developing a holistic model for digital library evaluation. J. Am. Soc. Inf. Sci. Technol. **61**(1), 88–110 (2010)
22. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Sølvberg, I.: Evaluation of digital libraries. Int. J. Digit. Libr. **8**(1), 21–38 (2007)
23. Tsakonas, G., Papatheodorou, C.: Analysing and evaluating usefulness and usability in electronic information services. J. Inf. Sci. **32**(5), 400–419 (2006)
24. Tsakonas, G., Kapidakis, S., Papatheodorou, C.: Evaluation of user interaction in digital libraries. In: Agosti, M., Fuhr, N. (eds.) DELOS Workshop on Evaluation of Digital Libraries, pp. 45–60. Padua (2004)

25. Steiner, C.M., Hillemann, E., Nussbaumer, A., Albert, D., Sweetnam, M., Hampson, C., Conlan, O.: The CULTURA evaluation model: An approach responding to the evaluation needs in an innovative research environment. In: S. Lawless, M. Agosti, P. Clough, O. Conlan (eds.) Proceedings of the First Workshop on Exploration, Navigation and Retrieval of Information in Cultural Heritage, SIGIR 2013, pp. 43–46. ACM, Dublin (2013)
26. Sweetnam, M., Agosti, M., Orio, N., Ponchia, C., Steiner, C., Hillemann, E.-C., Ó Siochrú, M., Lawless, S.: User needs for enhanced engagement with cultural heritage collections. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) Theory and Practice of Digital Libraries, TPDL 2012. LNCS vol. 7489, pp. 64–75. Springer, Berlin (2012)
27. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. Manag. Sci. **35**(8), 982–1003 (1989)
28. Gerdjikov, S., Mihov, S., Nenchev, V.: Extraction of spelling variations from language structure for noisy text correction. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. IEEE, pp. 324–328 (2013)
29. Steiner, C.M., Albert, D.: Tailor-made or unfledged? Evaluating the quality of adaptive eLearning. In: Psaromiligkos, Spyridakos, A., Retalis, S. (eds.) Evaluation in e-learning., pp. 111–143. Nova Science, New York (2012)
30. Brusilovsky, P., Karagiannidis, C., Sampson, D.: Layered evaluation of adaptive learning systems. Int. J. Contin. Eng. Educ. Life-Long Learn. **14**, 402–421 (2004)
31. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McCleeland, A.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor and Francis, London (1996)
32. Thong, J.Y.L., Hong, W., Tam, K.-Y.: Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences? Int. J. Hum.-Comput. Stud. **57**, 215–242 (2002)
33. Hampson, C., Lawless, S., Bailey, E., Steiner, C.M., Hillemann, E.-C., Conlan, O.: Metadata-enhanced exploration of digital cultural collections. Int. J. Metadata Semant. Ontol. **9**, 155–167 (2014)
34. Nussbaumer, A., Hillemann, E.-C., Steiner, C.M., Albert, D.: An evaluation system for digital libraries. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) Theory and practice of digital libraries. Second International Conference, TPDL 2012. LNCS, vol. 7489, pp. 414–419. Springer, Berlin (2012)