The CULTURA Evaluation Model: An Approach Responding to the Evaluation Needs of an Innovative Research Environment

Christina M. Steiner, Eva-C. Hillemann, Alexander Nussbaumer, Dietrich Albert
Knowledge Technologies Institute, Graz University of Technology
Inffeldgasse 13/V, 8010 Graz, Austria
{christina.steiner, eva.hillemann, alexander.nussbaumer, dietrich.albert}@tugraz.at

Mark S. Sweetnam
Department of History, Trinity College Dublin
Arts Building
Dublin 2, Ireland
sweetnam@tcd.ie

Cormac Hampson, Owen Conlan
KDEG, School of Computer Science and Statistics, Trinity College Dublin
O'Reilly Institute
Dublin 2, Ireland
{cormac.hampson, owen.conlan}@cs.tcd.ie

ABSTRACT
This paper presents the evaluation approach taken for an innovative research environment for digital cultural heritage collections in the CULTURA project. The integration of novel services of information retrieval to support exploration and (re)search of digital artefacts in this research environment, as well as the intended corpus agnosticism and diversity of target users posed additional challenges to evaluation. Starting from a methodology for evaluating digital libraries an evaluation model was established that captures the qualities specific to the objectives of the CULTURA environment, and that builds a common ground for empirical evaluations. A case study illustrates how the model was translated into a concrete evaluation procedure. The obtained outcomes indicate a positive user perception of the CULTURA environment and provide valuable information for further development.

Keywords
Cultural heritage, research environment, evaluation model, evaluation qualities, empirical study.

1. INTRODUCTION
Information retrieval technologies open up a range of possibilities for providing support in exploring, searching, and researching cultural heritage artefacts. Examples are automatic indexing and searching methods [9], normalisation of spelling variations in historical documents [10], extraction of entities representing persons, locations, events etc. from documents [3], the processing and mapping of dates and time intervals [8], or the application of social and influencer network analysis to digital collections. These new technologies and their integration and application in cultural heritage research environments and electronic information services require appropriate evaluation methodologies and thus pose specific requirements and challenges to evaluators. The present paper describes a comprehensive evaluation model accounting for that, which has been developed in the context of the CULTURA project[1]. The project aims at delivering a corpus agnostic research environment integrating a range of innovative services that guide, assist, and empower users' interaction with cultural heritage artefacts and take into account the diverse needs of different user groups. The novel methodological and technical approaches integrated, as well as the intended reusability of the technology with different collections and the diversity of users addressed as target audience challenges evaluation with respect to the use of sound and suitable methods across contrasting digital collections and diverse communities and users. Through the development of an evaluation model serving as a common ground for different evaluation studies, an appropriate level of comparability and generalizability of evaluation results can be maintained. The evaluation model is presented and a case study conducted with historians is outlined to illustrate how the investigation of the different evaluation axes and qualities of the model provide service-specific insights on the quality of the CULTURA environment and meaningful information for further development.

2. THE CULTURA PROJECT & SYSTEM
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 interested public for the first time. However, 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 artefacts. Furthermore, different types of users need 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 gain optimum engagement with these new digital humanities assets. Improved quality of access to cultural collections is a key objective of the CULTURA project [7]. 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 supports and features required for such a diverse range of users.

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

These services include recommenders, where links to relevant resources, based on the current document's entities (people, places etc.) and the user's overall interests, are displayed alongside the resource. Other features enabled by CULTURA include the creation of guided lessons. Importantly, CULTURA stores a detailed model of user actions within its environment, so by monitoring changes in this model it is possible to adapt the lesson to user interests, as well as improving recommendations. Annotations are another key service that CULTURA offers, and they can be used for private notes, group collaboration, or for teaching aids. All these features are offered by CULTURA on top of a keyword search facility, a text normaliser service, an entity based browser and social network visualisations, which provide complementary ways of exploring and understanding a cultural heritage collection. Due to the service-based architecture the suite of services can be extended iteratively over time, allowing new features to be offered to existing and future collections.

In order to validate the CULTURA environment, two major collections have been selected - the 1641 Depositions[2], held in Trinity College Dublin, Ireland and the IPSA Illuminated Manuscript Collection[3], which is distributed between a number of museums and universities around the world. In terms of the case study application presented in section 4, the instance of CULTURA with the 1641 Depositions collection was used.

3. THE CULTURA EVALUATION MODEL

CULTURA incorporates a range of different services, which necessitate specific consideration in evaluation to get comprehensive outcomes on the quality of the system. The interaction Triptych model [5][15] has been used as a starting point for a conceptual analysis of the components and aspects of CULTURA. This model distinguishes three main components: system, content, and user. Between these components the axes and qualities of evaluation can be identified: performance (system-content axis), usefulness (content-user axis), and usability (system-user axis). The model was extended for CULTURA (see Figure 1) to address the qualities specific to the research environment and its services and to form the theoretical basis for evaluation studies. In the following the evaluation qualities covered by the model are presented.

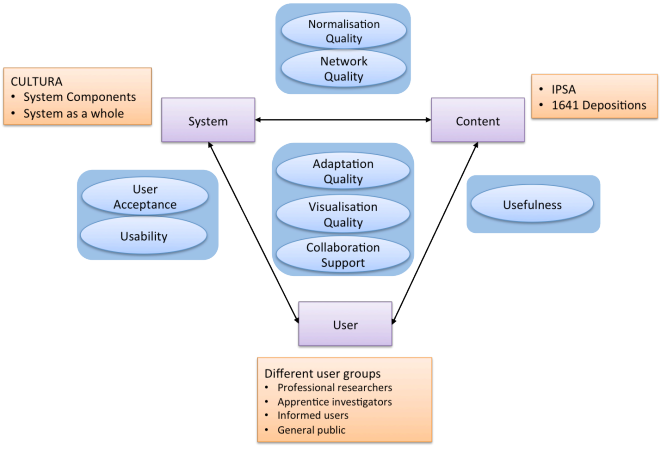

Figure 1. The CULTURA evaluation model.

Usefulness of content refers 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 personal user needs and/or the needs of the user group. A certain level of content usefulness is necessary for a meaningful evaluation of the other evaluation qualities.

Usability refers to the interaction 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. This evaluation quality is often considered using the ISO standard as a reference for collecting evaluation data (e.g. [6]).

User acceptance has been considered on the system-user axis of the evaluation model in addition to usability: Do users consider the research environment and its services acceptable? Users may not necessarily have a positive attitude towards the system, even if it is technologically sound. Commonly, the following user acceptance aspects are distinguished [4]: ease of use (related to usability aspects), usefulness (of the system – to be distinguished from usefulness of content), and behavioural intentions to use.

Adaptation quality refers to the interaction between system, content, and user: Is the adaptation provided by the CULTURA system appropriate and useful? This relates to users' perceived benefit of system adaptation/recommenders received (user-centred viewpoint) [13]. It can also be related to layered evaluation of adaptation [2], examining whether user variables are correctly inferred and whether adaptation decisions are appropriately taken.

Visualisation quality also refers to the interaction between all three components, system, content, and user: How do users feel about the visualisations provided by the research environment? In the context of CULTURA social and influencer network visualisations of collection contents and user communities are applied. Visualisation quality relates to users' perceived benefit of the visualisations provided (helpfulness, insights gained etc.).

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

Performance: *Normalisation quality and network quality*. The aspect of performance (system-content axis) is usually not directly visible to the users and often difficult to evaluate via user feedback. In CULTURA the performance aspect is operationalized as normalisation and network quality. Normalisation quality refers to text normalisation as well as entity extraction from text, i.e. to the quality and accuracy of the output of these processes. Network quality refers to whether relations between the entities of digital artefacts are accurately technically presented in the visualisations. Network quality thus investigates the accuracy of the data visualisations and the occurrence of inconsistencies between the entity data and the visualisation.

4. EMPRICAL CASE STUDY

To demonstrate the applicability and application of the evaluation model, this section presents a small-scale evaluation of the CULTURA environment for the 1641 Depositions collection with professional researchers. The study used an evaluation method that was set up in alignment with the evaluation model.

[2] http://1641.tcd.ie/
[3] http://www.ipsa-project.org/

4.1 Method

Evaluation instruments defined in line with the evaluation model were: an online survey covering items or scales on all evaluation qualities, semi-structured interviews, as well interaction logs as quantitative data complementing participants' self-reports. Thirteen professional researchers in history took part in the study. Log data was available for the whole sample, while only seven persons (6 male, 1 female) completed the survey. Participants were on average 39 years old, with a range from 28 to 47 years.

Participants were introduced to the CULTURA environment and its functionality. Subsequently, they had the possibility to use the system in their own time. Interaction data was recorded for the whole duration of usage. Users visited on average 15 pages while interacting with the system. Noticeably, the users were unlikely to 'play' with the system, but tended to ask for demonstrations of the different services. After working with the system, participants completed the online survey and took part in an interview.

4.2 Results

The *usefulness of content*, i.e. of the 1641 Depositions collection, was assessed very high, with $M = 6.64$ ($SD = 0.94$) on a scale ranging from 1-7, as it would be expectable for a user group with explicit expertise and interest in the digital collection in question.

The standard *usability assessment* [1] yielded an average score of 68.21 ($SD = 19.18$), indicating good usability (possible score range 0-100). Participants did not have a highly consistent perception of the system's usability, though, which might be due to a variable level of comfort with technology, in general, as it could be identified in the interviews. A number of participants highlighted the need for a guided tour introducing the system's features and how to use them.

User acceptance, assessed with an instrument adopted from prior research [14], was positive on all aspects (on a 1-7 scale, in each case), with the best result for behaviour intention ($M = 6.0$, $SD = 1.83$), arguing for participants' willingness and interest in actually using the system. The perceived usefulness of the CULTURA environment was also very good ($M = 5.89$, $SD = 1.81$), the score on ease of use ($M = 5.11$, $SD = 1.88$) was good.

For *adaptation quality* and *visualisation quality* subscores on estimated usage, usability, and perceived benefit, as well overall scores were calculated (see Figure 2). As can be seen a similar pattern can be identified for both qualities, with visualisations scoring generally slightly better than recommenders. Users indicated rather scarce usage of the recommenders and visualisations (recommenders $M = 3.12$, $SD = 2.04$; visualisations $M = 3.57$, $SD = 2.15$). Log data reinforces this finding: while 5 people did not use the visualisation service at all, the other 8 users visited 1 to on maximum 5 visualisations ($M = 1.46$, $SD = 1.56$ for $N = 13$); recommendations were on average used only 0.46 ($SD = 1.66$) times, and part of the users completely waived them. Overall quality scores, as well as usability and perceived benefit scores were assessed with medium quality in both cases, except for the benefit of visualisations, which scored more positively. Confirming this, in the interviews researchers expressed an appreciation of the possibilities offered for new insights into the Depositions by visualisations, but also pointed to the need for more flexible visualisations. Interviews also confirmed the usefulness of recommendations for exploring the collection – especially when not intimately familiar with the content, which explains why participants did not extensively use them themselves. In addition, users stressed the need for transparency – they wanted to know not only what is recommended, but also why they are seeing a particular recommendation.

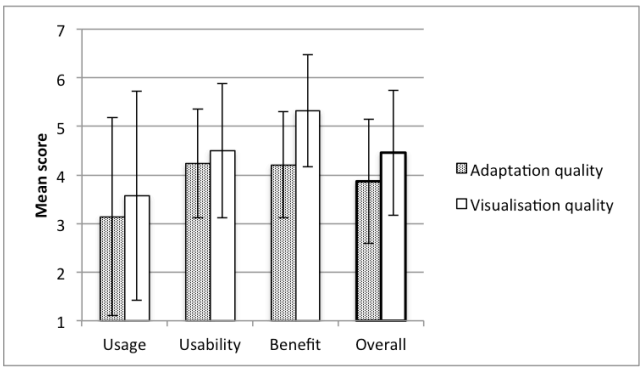

Figure 2: Results (mean scores with SD) on adaptation quality and visualisation quality.

Collaboration support was perceived as good ($M = 5.36$, $SD = 0.99$); researchers felt the system may assist them in collaborating with others. The assessment of the annotation service, which goes beyond pure collaboration features, was even better ($M = 6.29$, $SD = 1.25$). However, although users' were amenable to the idea of annotations, they did not take or share any annotations themselves.

Responses on normalised search were very positive, thus indicating an excellent user-centred assessment of *normalisation quality* ($M = 6.64$, $SD = 0.64$). Open comments highlighted that this feature is highly valuable and the majority of users had also made use this feature during the trial. Interviews, however, uncovered that users had concerns with respect to the accuracy of information extraction. In the lists of automatically extracted entities that appeared alongside the transcription of each deposition, the occurrence of errors – even if they were isolated – tended to have a devastating effect on users' confidence in the system. Addressing these concerns was of importance to their adoption of the research environment.

5. CONCLUSION

This paper introduced the evaluation approach developed and applied in the CULTURA project to respond to the evaluation needs of an innovative research environment for digital cultural heritage collections. Through the alignment of all evaluation tasks to the common evaluation model, general comparability of results is maintained. This is especially important since the CULTURA system is intended as a corpus agnostic research environment usable with different digital collections and addressing a broad range of user groups along the dimension of expertise. The final aim for evaluation is therefore to prove the benefit of the CULTURA environment independent of a specific collection and type of user. A comparison and consolidation of results over user groups and collections allows finding out about the benefits and issues that are of general interest and the overall quality of the research environment and its integrated services.

Evaluation studies other than the one presented herein have involved task-based evaluations (e.g. [11]), where users are requested to work on a predefined task while trialling the system. Such a task-based procedure enables a more detailed investigation of evaluation qualities, like adaptation quality or collaboration support, by drawing conclusions from the actual use of the system. This kind of procedure was unsuitable in the present case,

though, since professional researchers wished to explore the system themselves without being forced to work on a given task. Moreover, other user studies on the CULTURA environment have also involved the comparison with the original web application of the respective digital collection or with the baseline version of the CULTURA system, without intelligent services.

Although participants in our case study acknowledged the general usefulness of the adaptive recommenders, their scarce actual usage highlights another important issue: the recommender service is rather intended for users with no or low prior knowledge in the collection than for expert researchers, who do not need or even do not want to have any guidance. This points up that the qualities of the evaluation model need to be investigated with appropriate groups of users, corresponding to the target audience of the services underlying this quality.

The results obtained from the presented study were consolidated with results obtained from other and larger scale user trials to derive implications for further development. Changes already implemented in the meantime were appreciated in more recent user trials and were singled out as being especially valuable in terms of building users' comfort with and confidence in the CULTURA environment.

The presented evaluation model is considered to have high potential for reuse in other research environments. It provides a valuable starting point for identifying the axes and topics of interest in other evaluation contexts and for specifying the actual evaluation design and evaluation instruments to be applied. The evaluation model is also used as a basis for the development of an evaluation service in the CULTURA project [12], aiming at supporting evaluators in planning, carrying out, and analysing evaluations. Through explicitly specifying the quality model underlying an evaluation, data collection can be systematized and automated reports based on the mapping to evaluation qualities can be derived. Future work will focus on triangulating data gathered via different modes (i.e. explicit retrospective or on-line user feedback, and non-invasive log and sensor data) by using the evaluation model as a reference base.

6. 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 could not be realised without the close collaboration between all CULTURA partners.

7. REFERENCES

[1] Brooke, J. 1996. SUS: a „quick and dirty" usability scale. In P.W. Jordan, B. Thomas, B.A. Weerdmeester & a. L. McClelland (Eds.) *Usability evaluation in industry* (pp. 189-194). Taylor & Francis, London.

[2] Brusilovsky P., Karagiannidis C., and Sampson D. 2004. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning* 14, 402-421.

[3] Carmel, D., Zwerdling, N., and Yogev, S. 2012. Entity oriented search and exploration for cultural heritage collections. *World Wide Web 2012 European project track. Lion, France.*

[4] Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management Science* 35(8), 982-1003.

[5] Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, Peters, C., and Solvberg, I. 2007. Evaluation of digital libraries. *International Journal on Digital Libraries* 8, 21-38.

[6] Gediga, G., Hamborg, .-C., and Düntsch, I. 1999. The IsoMetrics usability inventory: An operationalisation of ISO 9242-10. *Behaviour and Information Technology* 18, 151-164.

[7] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. 2012. The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In *Proceedings of the 4th International Euromed Conference, Limassol, Cyprus* (pp. 668-675). Springer, Heidelberg.

[8] Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvönen, E., and Bandini, S. 2010. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *International Journal of Human-Computer Studies* 68, 549-560.

[9] Koolen, M., Kamps, J., and Keijzer, V.d. 2009. Information retrieval in cultural heritage. *Interdisciplinary Science Reviews* 2-3, 268-284.

[10] Lawless, S., Hampson, C., Mitankin, P., and Gerdjikov, S. 2013. *Normalisation in historical text collections*. Accepted for publication at Digital Humanities, University of Nebraska-Lincoln.

[11] Lin, J. 2007. Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting* (pp. 212–219). Association for Computational Linguistics, Rochester.

[12] Nussbaumer, A., Hillemann, E.-C., Steiner, C.M., and Albert, D. 2012. An evaluation system for digital libraries. In Zaphiris, P., Buchanan, G., Rasmussen, E., and Loizides, F. (Eds.), *Theory and practice of digital libraries. Second International Conference, TPDL 2012. LNCS vol. 7489* (pp. 414-419). Springer, Berlin.

[13] Steiner, C.M. and Albert, D. (2012). Tailor-made or unfledged? Evaluating the quality of adaptive eLearning. In Psaromiligkos, A. Spyridakos, & S. Retalis (Eds.), *Evaluation in e-learning* (pp. 111-143). Nova Science, New York.

[14] Thong, J.Y.L., Hong, W., and Tam, K.-Y. 2002. Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences? *International Journal of Human-Computer Studies* 57, 215-242.

[15] Tsakonas, G. and Papatheodorou, G. 2006. Analysing and evaluating usefulness and usability in electronic information services. *Journal of Information Science* 32, 400-419