Metadata-enhanced exploration of digital cultural collections

Article in International Journal of Metadata Semantics and Ontologies - April 2014
DOI: 10.1504/IJMSO.2014.060342

6 authors, including:

Cormac Hampson
Trinity College Dublin
35 PUBLICATIONS 155 CITATIONS
SEE PROFILE

Séamus Lawless
Trinity College Dublin
64 PUBLICATIONS 573 CITATIONS
SEE PROFILE

Eoin Bailey
Trinity College Dublin
12 PUBLICATIONS 54 CITATIONS
SEE PROFILE

Eva Hillemann
Graz University of Technology
18 PUBLICATIONS 91 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Cultura View project

Project BIDAYS View project
Metadata-enhanced exploration of digital cultural collections

Cormac Hampson*, Séamus Lawless and Eoin Bailey

Knowledge and Data Engineering Group,
Trinity College Dublin,
Ireland
Email: cormac.hampson@cs.tcd.ie
Email: seamus.lawless@cs.tcd.ie
Email: eoin.bailey@cs.tcd.ie
*Corresponding author

Christina M. Steiner and Eva-Catherine Hillemann

Knowledge Technologies Institute,
Graz University of Technology,
Austria
Email: christina.steiner@tugraz.at
Email: eva.hillemann@tugraz.at

Owen Conlan

Knowledge and Data Engineering Group,
Trinity College Dublin,
Ireland
Email: owen.conlan@cs.tcd.ie

Abstract: The increased digitisation of cultural collections and their availability on the World Wide Web has made access to these valuable documents much easier than ever before. However, despite the increased availability of access to cultural archives, curators still struggle to instigate and enhance engagement with these resources. The CULTURA project is actively addressing this issue through the development of a metadata-driven personalisation environment for exploring cultural collections and instigating collaborations. This paper discusses the state of the art in this area and the various innovative approaches used in the CULTURA project, with a special focus on how the underlying metadata helps to facilitate its semantically rich environment. An evaluation of the CULTURA project with students is detailed, highlighting its relevance for those without domain expertise. This is vital as the system needs to adapt and support a full spectrum of users, from professional researchers to interested members of the public.

Keywords: CULTURA; metadata; exploration; personalisation; digital humanities; normalisation; social network analysis; entity extraction; annotation.


Biographical notes: Cormac Hampson is a post-doctoral researcher in the School of Computer Science and Statistics in Trinity College Dublin. He has a BSc in Multimedia from Dublin City University and an MSc in Ubiquitous Computing from Trinity College Dublin. His PhD detailed an expert-supported approach to data exploration, and he has been involved in several European projects in the cultural heritage and eLearning domains. His research areas include data exploration, personalisation and digital humanities.

Séamus Lawless is an Assistant Professor in the School of Computer Science and Statistics in Trinity College Dublin. His research interests are in information retrieval, information management and digital humanities with a particular focus on adaptivity and personalisation. The common focus of this research is the application of technology to support enhanced, personalised access to knowledge. He has over 40 publications in major journals, ACM, IEEE and other conferences. He is a funded investigator in CNGL, an SFI-funded Centre for Science, Engineering and Technology (CSET). This includes leading the research agenda of the Search and Discovery theme. He is also a lead investigator in the CULTURA EU FP7 project.
Eoin Bailey graduated from Trinity College Dublin, Ireland in 2003 with a degree in Electronic and Computer Engineering. In 2004 he was awarded an MSc in Networks and Distributed Systems from Trinity College Dublin, Ireland. His research areas of interest include geographic information systems, applying digital technologies to novel domains, and personalisation systems.

Christina M. Steiner completed her Diploma in Psychology in 2004 and received her Dr.rer.nat in 2012 at the University of Graz, Austria. From 2004 to 2011 she has been a researcher at the Department of Psychology of the University of Graz. Since 2010 she has been working at the Knowledge Technologies Institute of Graz University of Technology. She has been involved in several European projects on technology-enhanced learning, and on digital libraries and cultural heritage. Her research interests include knowledge and competence representation and structuring, psycho-pedagogical frameworks for adaptation, user-centred evaluation, as well as concept mapping.

Eva-Catherine Hillemann has graduated from the University of Graz, Austria with a degree in psychology. From 2008 to 2011 she worked as researcher at the Department of Psychology of the University of Graz. Since 2011, she has been working at the Knowledge Technologies Institute of Graz University of Technology. Her current research focuses on the evaluation of technology-enhanced learning and research environments as well as on psycho-pedagogical aspects in e-learning, e-education, and e-assessment. Additionally, she is currently writing her PhD thesis in which she is investigating how Open Learner Models influence and affect learner’s meta-cognitive abilities in the learning process.

Owen Conlan is a member of the School of Computer Science and Statistics at Trinity College Dublin. He leads research in the areas of personalisation and adaptation in the Centre for Global Intelligent Content and is the coordinator of the FP7-funded CULTURA project. He is well known for his research in personalisation and visualisation as applied to the domains of Digital Humanities and Technology Enhanced Learning. As part of his doctoral studies he proposed the innovative narrative approach to personalisation and it is this approach that has underpinned advancements such as personalised exploration and non-invasive personalisation.

This paper is a revised and expanded version of a paper entitled ‘CULTURA: a metadata-rich environment to support the enhanced interrogation of cultural collection’ presented at the ‘6th Metadata and Semantics Research Conference (MTSR 2012)’; 28–30 November 2012, Càdiz, Spain.

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 interested public for the first time. However, the full value of these heritage treasures is not being realised. 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. As a result, they often fail to attract and sustain broad user engagement leading to limited communities of interest. Thus, there still remain important challenges in the presentation of new digital humanities artefacts to the end user.

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, especially those collections which are not exhibited physically, is a key objective of the CULTURA project (Hampson et al., 2012; Bailey et al., 2012). 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 (Sweetnam et al., 2013). To this end, CULTURA is delivering a corpus agnostic environment, with a suite of services to provide the necessary supports and features required for such a diverse range of users. A prototype of the CULTURA system is available online at http://cultura-project.eu/

A central aspect of this environment is its use of rich metadata (user generated, computer generated and expert generated) which has been generated or enhanced through natural language processing, entity extraction and Social Network Analysis (SNA) techniques, in order to support collaborative exploration, interrogation and interpretation of the underlying cultural resources. In this paper, Section 2 discusses some related work in the field of digital humanities, Section 3 outlines the key challenges of this research. Section 4 introduces two cases, studies which are central to the CULTURA project; Section 5 discusses the various components and features of CULTURA’s architecture; Section 6 details an evaluation study of the current CULTURA prototype with school students; and Section 7 summarises the paper and discusses the future work that remains to be undertaken.
2 Related work

There has been substantial effort in the area of digitisation and cultural heritage preservation. Much of this work has, until recently, been focused on the creation of digital representations of cultural artefacts and the creation of metadata and documentation associated with this. The result of this effort is that there is a vast collection of content available to digital humanists in the form of text, images and other representations.

Textual content resources include collections arranged by theme, such as the Biodiversity Heritage Library\(^1\) and from institutional collections, such as the Bayerische StaatsBibliothek\(^2\) digital collections library. These collections include varying levels of metadata and some include detailed pictures which are associated with the text. It is important to note that in many cases, textual content actually refers to a complex cultural artefact that includes an image of the original manuscript, transcribed text associated with the content, and metadata which can describe the content, the nature of the document and the provenance of the digital artefact. Another important type of artefact is collections of more visual material: images, illuminated manuscripts, paintings, etc., with detailed metadata records associated with each image. A persistent challenge exists in supporting content providers in adding such complex cultural artefacts to federated content/metadata repositories, such as Europeana.\(^3\) The DM2E (Digital Manuscripts to Europeana) project\(^4\) addresses this challenge by providing two tools – Korbo and Pundit. Korbo is an aggregation platform that supports users in gathering baskets of cultural resources, while Pundit is a semantic annotation tool to enable users to annotate resources. Tools such as these provide mechanisms to ease the introduction of content into repositories.

While there have been recent attempts to use adaptive hypermedia techniques to support the personalised retrieval, interrogation and presentation of cultural heritage content collections, these have to date been limited. The MultimediaN9C Eculture project\(^5\) aims to provide multimedia access to distributed collections of cultural heritage objects. The aim of the project is to support the generation of various types of personalised and context-dependent presentations of cultural material. However, the current system only provides static semantic search across entities in manually annotated content collections. The CHIP project\(^6\) aims to provide personalised presentation and navigation of the Rijksmuseum cultural resources. The artwork recommender supports the rating of artworks/topics to generate a user profile, which is then used to drive future artwork recommendations. The ‘Tour Wizard’ is a web-based tool which uses the user profile to semi-automatically generate personalised museum tours. In the MOSAICA project\(^7\), a mobile device-based demonstration is used to engage novice and intermediate users. The system does provide virtual visitors with access to structured descriptions of collections through a search interface, but little adaptivity or personalisation of the experience is used.

The QViz\(^8\) project has some similarities in approach to the CULTURA project, in that it makes explicit recognition of the value of users as members of communities and as contributors to digital cultural heritage collections. The focus of the QViz system is on temporal and spatial search and retrieval of archival content. While QViz is a social semantic application, facilitating user contribution and structured representation of knowledge, it does not have a personalised or adaptive aspect. Because CULTURA is producing a generalisable solution, it must be able to add value to a wide range of digital cultural heritage collections, of which there are many. One example is the Europeana project,\(^9\) which represents metadata from collections across many EU member states. While Europeana does not directly host content, it is a large repository of metadata which could be processed, alongside a specific collection’s content, to seed the CULTURA environment. Many projects within the cultural heritage domain, including PATHS\(^10\) and Natural Europe,\(^11\) already encompass rich metadata from Europeana within their environments.

3 Challenges in the digital humanities

The rise of ‘i’ and ‘my’ as prefixes for various web portals (e.g. iTunes\(^12\) and MySpace\(^13\)) and web services (e.g. iCloud\(^14\)) is intended to give the impression of personal tailoring of content and service to an individual user to enhance that individual’s experience. Typically, however, such services tend to focus on (a) identification and ranking of relevant content or services, (b) simplistic ‘personalisation’ of the content presentation by inclusion of the user’s name, recently used resources etc. or (c) simple augmentation of screen layout.

To effectively empower communities of researchers with personalised mechanisms, which support the collaborative exploration, interrogation and interpretation of complex digital cultural artefacts, the adaptivity provided in CULTURA must be more integrated and intelligent than in the portals described above. Such next generation adaptivity, as espoused by CULTURA, must support the dynamic composition and presentation of digital cultural heritage resources. However, just automated adaptivity is not enough. Ensuring that the user is in control of the personalisation process is essential. For example, as part of the CULTURA environment, researchers are always offered access to a visualisation of their user model (displayed as a tag cloud) that they can interact with. This empowers the user to both reflect on their activities and to tune the model to meet their needs. The model is continuously added to by introducing entities (people, places, events, etc.) related to the resources that the users work with. These entities are added with varying degrees of confidence depending on the form of interaction the user has with the resource. The user model enables both personalised entity-level recommendations of appropriate content and suggestions for engagements with other users. As described above, user control is a key to ensuring such recommendations are appropriate.

A common challenge in the humanities is that historical language hinders the accessibility of historical text documents. One solution to this problem is the use of a computational
historical lexicon, supplemented by computational tools and linguistic models of variation. However, because of the absence of language standards, multiple orthographic variations of a given word or expression can be found in a collection of material, even in the same document. Hence, issues arising from the need to contend with noisy inputs, the impact noise can have on downstream applications, and the demands that noisy information places on document analysis are addressed by CULTURA. This challenge is addressed by both statistical and rule-based normalisation approaches within the project. This hybrid approach, tested on the 1641 Depositions corpus, has resulted in a highly accurate normalisation (~95%) (Lawless et al., 2013).

SNA can be used to analyse the people and relationships contained within humanities content collections. However, the effective application of SNA techniques to content, which has major inconsistencies in the naming and identification of entities, poses a significant challenge which must be overcome. Importantly, CULTURA applies normalisation techniques to the entity extraction process (see Section 5.3) which improves the effectiveness of SNA greatly. Furthermore, curators of the digital artefacts are able to visualise the entity graph (exported after the entity extraction process occurs) using PreMapper (Hampson et al., 2013). Fundamentally, the PreMapper tool supports curators of data to easily add, delete, merge, disambiguate and edit entities using a GUI. Typically digital cultural heritage collections contain complex relationships between entities which must be identified and extracted from the artefacts. This is an area CULTURA also directly addresses by augmenting the existing metadata with new attributes. A second challenge for SNA is to leverage the user communities, activities, contributions and profiles to discover the rich influence network that interlinks users of these digital humanities content collections. The application of SNA to cultural artefacts as well as to the community that surrounds those artefacts is novel in the digital humanities.

4 Case studies

In order to validate the CULTURA environment, two major artefacts have been selected – the 1641 Depositions, held in Trinity College Dublin, Ireland (not on public display) and the IPSA Illuminated Manuscript Collection which is distributed between a number of museums and universities around the world. Improving the quality of access to cultural collections, especially those collections which are not exhibited physically, is a key objective of the CULTURA project; hence, the selection of these important works is aptness. These resources and the communities of users who work with them are central to the design, development and evaluation of the CULTURA environment. Each is now discussed in turn.

4.1 The 1641 Depositions

The 1641 Depositions are 17th-century manuscripts that comprise about 8000 witness statements, examinations and associated materials, in which Protestant men and women of all classes and from all over Ireland told their experiences following the outbreak of the rebellion by the Catholic Irish in October 1641. This body of material is unparalleled anywhere else in early modern Europe and provides a unique source of information for the causes and events surrounding the 1641 rebellion and for the social, economic, cultural, religious and political history of 17th-century Ireland, England and Scotland.

The 1641 Depositions have been digitised and transcribed and are being used to validate the techniques implemented in CULTURA. From a technological perspective, the 1641 Depositions represent a textually rich digital humanities collection, which is characterised by noisy text, inconsistent sentence structure, grammar and spelling. The Early Modern English language manuscripts contain rich metadata and descriptions of individuals, locations, events, social structures and contrasting/conflicting narratives. These artefacts have active communities of interest because of their wider social and historical implications that transcend geographical and chronological boundaries and continue to shape opinions and values to this day. The 1641 Depositions represent an ideal example of a digital humanities collection, which has deep resonance with social and cultural issues encountered throughout Europe.

4.2 The Imaginum Patavinae Scientiae Archivum

The Imaginum Patavinae Scientiae Archivum (IPSA) collection is an archive of illuminated medieval astrological and herbal manuscript codices dating from the 14th century with Latin, Paduan and Italian language commentaries. Herbals are manuscripts which contain hand-drawn depictions of plants, such as trees, bushes or shrubs and their parts, such as flowers or leaves. The IPSA collection contains manuscripts written and illustrated by the Paduan School, and successive manuscripts produced in Europe under its influence. Such manuscripts have the rare characteristic of containing high quality and very realistic illustrations. IPSA is a combination of digitised images of the manuscripts and related metadata descriptions.

From a technical perspective, IPSA represents a very different kind of digital humanities collection to the 1641 Depositions collection. The IPSA collection is primarily image based, with substantive metadata available. This metadata not only provides descriptive passages, but is also historically valuable as it captures the scientific processes which were prevalent during the creation of the original collection. However, the IPSA metadata is user generated which can lead to inconsistencies in terminology, spelling and grammar. The metadata contains descriptions of entities, individuals, activities and locations in multiple languages. 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.
5 The CULTURA architecture

CULTURA consists of multiple distinct services all accessed via the CULTURA portal. The services available in CULTURA are shown in Figure 1 and include personalised search tools, faceted search tools, annotators, social network exploration tools and recommenders. One of the key challenges for the CULTURA architecture is to reconcile the various metadata models at runtime, in order to seamlessly provide the end user with the most appropriate content and services. As seen in the data layer of Figure 1 the metadata includes:

- User models and logs that contain data on user preferences and their previous interaction with the portal (Hampson et al., 2012). The user models are stored in a MongoDB database as this allows the schema to evolve easily over time when new services are added.
- Content metadata that describes the digital resources, such as the 1641 Depositions, which were encoded according to the text encoding initiative.
- Entity metadata extracted using named entity recognition (Carmel et al., 2012) and stored as XML.

Each individual model contains a shared identifier for every artefact in the collection to facilitate the correlation of data from different models, but the reconciliation of these models at runtime is a dynamic process that does not involve the production of a new static model. An example of such a process is when a service is triggered by a user’s interactions with the CULTURA portal, with requests sent from the presentation layer to the service via its API. For instance, when a person is looking at one of the 1641 Depositions, entities from that document (people, places, etc., stored as metadata in the data layer) are extracted and cross-referenced with the entities stored in the relevant user model. Recommended depictions, based on the entities in the current content and those stored in the user model, are then calculated in the control layer by the recommender widgets. These recommendations are then rendered in the presentation layer for the user to view (see Figure 2).

The CULTURA portal utilises Drupal as it provides numerous services that, while essential 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 functionality. 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. SNA of people mentioned within the 1641 Depositions can be displayed as a ‘wheel’ as shown in Figure 3.

5.1 Normalisation of cultural collections

Performing document analysis techniques (i.e. information extraction) on historical texts, which contain non-standard spelling, historical grammar and many old word forms, is a non-trivial challenge requiring normalisation of word spelling and entity extraction. The primary purpose of the normalisation process is to produce documents without historical variations on letter level. Within CULTURA, this normalisation process enables better identification of entities, e.g. people, places, events, dates, as well as facilitating improved search across the collection by taking account of spelling variants of a search term (Lawless et al., 2013). This is achieved by producing XML metadata for each resource, which describes the mapping and offset of each normalised term in the document. The statistical model built in CULTURA to automatically normalise historical texts utilised a training set of manually normalised documents from the 1641 Depositions. These manually normalised documents were randomly selected and accounted for approximately 6% of documents from the collection. The translation model was developed on top of the previously developed OCR correction methodology (Mihov et al., 2007) and importantly can be applied to other resources besides the 1641 Depositions (Gerdjikov et al., 2013).

Figure 1 The CULTURA architecture (see online version for colours)
Effective data modelling is one of the crucial aspects of supporting entity-oriented search for data-centric systems. In the context of digital humanities, this data modelling challenge has two specific characteristics. The first requirement is to allow the incorporation of new concepts which augment the original data during the research process and are not explicitly encoded a priori, e.g. detection of entities and events within a document. The other requirement is to support a layer of services that allow a range of user’s actions. These actions typically include the manual manipulation of existing data, the user referencing of specific data elements or the interaction between sets of users.

In both cases, the system schema must evolve over time to reflect the work of researchers and others. Therefore, the data management part of the system must support easy on-the-fly modifications of the underlying schema. This requirement rules out relational databases, since modifying the columns of a populated relational database table is a costly task. However, conceptual modelling based on the
entity-relationship model, which is commonly used in the process of relational databases schema definition, is an effective methodology for capturing data requirements. Hence, this approach has been chosen for the entity-oriented search within CULTURA; and entity-relationship schemas have been manually defined for different digital cultural archives using its environment, e.g. the IPSA collection and the 1641 Depositions.

Once an entity-relationship schema has been generated, then entity-relationship data is indexed (see Section 5.3) which populates the model and encodes it as an XML. The model allows expressive search capabilities ranging from simple keyword search to complex structured queries (Yogeved et al., 2012). The 1641 Depositions data has already been modelled and processed as part of the entity-oriented search component (Carmel et al., 2012), providing a powerful exploratory search system. This entity-oriented component provides CULTURA with a powerful tool for incremental research, where results of prior analysis are accessible for search and exploration. Furthermore, it enables the improvement of existing analysis processes and supports the discovery of new insights into the collection to be discovered. Importantly, apart from entity-oriented search, this same model is also used to help power SNA (see Section 5.3) and content recommendations (see Figure 2). Finally, the entity-oriented approach also facilitates researchers and other interested users to be added as entities in the system and link them to existing entities. Such connections between the system’s users and the data can afterward be used to automatically define the social network of the community interested in the digital archive, and later this social network can be used to provide social-based services to the community.

5.3 Network analysis of cultural collections

Normalised historical texts can be analysed using Influencer Network Analysis (INA), a form of SNA, which is used to identify the social structure described in the historical texts and the influential people involved in the incidents portrayed. Influential entities not only include the individuals involved within the collection, but also the issues, topics and opinions that are detailed. Entity extraction is performed on the historical texts using a UIMA pipeline and IBM’s LanguageWare platform. Lexicons are constructed manually and enhanced through the normalisation process. These lexicons are then ingested by the UIMA pipeline and used to identify the entities contained within the texts (see Figure 4).

Relationships between entities are also identified and can augment the existing metadata from the collection. These entities and relationships are graphed and shown to users allowing a user to explore the relationships between entities across multiple historical texts (see Figure 3). As mentioned in Section 3, the PreMapper tool (Hampson et al., 2013b) enables curators to view a graph of extracted entities and to add, delete, merge, disambiguate and edit these entities using a GUI.

Figure 4 The normalisation and entity extraction process flow (see online version for colours)
Network analysis also enables the tracking of entities both temporally and geographically. This is achieved using algorithms developed by Commetric EEOD. Owing to the scale and complexity of the data typically held in digital cultural collections, this has rarely been attempted by humanities researchers. Furthermore, apart from performing network analysis on the historical texts themselves, the same technique can be used to analyse how a community of users engages with the collection interact. This process makes it easier for users to discover relevant experts, for new communities to be created based on similar activities, and for community collaboration and contribution to occur.

5.4 Personalisation and adaption techniques in CULTURA

A core element of the CULTURA architecture is its use of personalisation and adaption techniques. These techniques have been heavily influenced by Adaptive Hypermedia (AH) and adaptive web systems research, which are concerned with improving the retrieval and composition of information. This improvement is achieved by creating a more context-sensitive and personalised interaction with digital content and is often predicated on rich metadata (Brusilovsky et al., 2007).

One reason why novice users struggle to engage with large cultural collections is a lack of guidance when they initially encounter the set of resources. To counteract this, CULTURA employs a four-phase personalisation approach (see Figure 5). Users with little experience of the underlying resources typically start their investigations within the guide phase. Here, CULTURA employs a ‘narrative’ module, which enables resources within the collection to be sequenced on a specific theme (e.g. the importance of books in 17th century Ireland as described in Section 6). Furthermore, how these resources are rendered to the user (text, visualisation etc.) can also be specified within the narrative metadata (encoded as XML). This process is especially useful in providing users with a path through specific content, though it does not limit their ability to use these sequenced narratives as a springboard for their own investigations. Within the four-phase personalisation approach, this involves stepping from the guide phase to the explore phase (number 1 in Figure 5).

In the explore phase, CULTURA offers tool assistance (e.g. data enriched maps, entity based search, SNA) to support exploration and browsing of the underlying resources. At any stage a user can return to where they left their path in the guide phase, and users with little prior knowledge of the resources often flick between explore and guide phases several times. In contrast, professional researchers with a deep understanding of the collection typically spend the majority of their time within the explore phase and may never involve themselves with the guide phase. Importantly, by monitoring the user model metadata as they explore the resources, the narrative path itself can be adapted. This is achieved by selecting documents for their path that most closely match the user’s interests, and can result in the path being enriched with further resources and concepts.

Within the guide or explore phase, a user will be given personalised suggestions for related content (see Figure 2) or tools to view resources in. This process occurs in the suggest phase (which works in parallel with the guide and explore phases), with hints pushed to users for their review (number 2 in Figure 5). These hints are influenced by the content the user is currently viewing, as well as the data stored in their user model (e.g. search terms, entities commonly viewed and annotations created). At any stage within the guide or explore phases, a user may enter the reflect phase which involves viewing their user model (rendered as a tag cloud) and seeing what interests the CULTURA environment has associated with them (number 3 in Figure 4). The reflect phase is not static, and users can delete metadata terms in their user model or manually reduce their size (thus reducing their degree of influence). Importantly, any changes made during the reflect phase directly impacts on the suggest phase and the recommendations that eventually filter down into the guide and explore phases (number 4 in Figure 5). By espousing the four-phase personalisation approach, CULTURA dynamically adapts to users, and renders useful suggestions to them at appropriate times. Moreover, this process provides mechanisms that are appropriate for a range of users with different levels of ability or different interests in the underlying resources.

The application of personalisation techniques helps empower experienced researchers, novice researchers and the wider community to discover, interrogate and analyse cultural heritage resources. Figure 2 shows one example of the recommended content shown to users who browse the 1641 Depositions using CULTURA. When a user views a deposition, entities (people, places etc.) are extracted from the text, and complementary depositions that also mention these entities are located within the collection. In Figure 2, blue text links to depositions related to the entities ‘Trim’ and ‘Lismore’ are displayed. This box is rendered beside the deposition text and enables users to quickly locate new resources that are relevant and to also know why these resources are being recommended to them. Furthermore, by
clicking on one of these links, the user model is updated with the corresponding entity, as it indicates a user interest in said entity.

CULTURA employs MongoDB\textsuperscript{4}, a scalable high-performance NoSQL database, to store its user models. One of the main benefits of MongoDB is that it allows the database schema to evolve over time, which is very useful in a service-orientated environment such as CULTURA, where new suites of tools are being introduced gradually. Complementary to the user models stored in MongoDB are the detailed logs saved by the CULTURA environment. These provide further rich metadata that can be exploited by the system for personalisation and adaptation purposes.

5.5 Annotation service

Content Annotation Tool (CAT) (Munnelly et al., 2013; Ferro et al., 2013) is a web annotation tool that enables users to manually annotate both text and images within the CULTURA environment. The current granularity for annotation of text is at the level of the letter; for image annotations the granularity is at the level of the pixel. This allows for extremely precise document annotation, which is very relevant to the digital humanities domain, due to the variety of different assets that prevail. Further to allowing a user to comment on document text and images, the annotations created using CAT allows a user to link their annotations to other resources internal and external to the CULTURA environment. This enables an annotation to become an edge between nodes (resources) in a unidirectional graph.

The annotations created by CAT (see Figure 6) are themselves stored in Flexible Annotation Semantic Tool (FAST) (Ferro, 2009), an annotation service developed by Padua University. FAST is a flexible, semantically linked data store for annotations. It is capable of managing both structured and unstructured contents from its users. The FAST annotation service adopts and implements the formal model for annotations proposed by Agosti and Ferro (2008), which has also been also embedded in the reference model for digital libraries developed by DELOS, the European network of excellence on digital libraries (Candela et al., 2007).

By default, the scope of an annotation is set to private, meaning that only the person who created the annotation can see it. These annotations may serve as reminders or notes within the document for the user. Alternatively, the scope of an annotation can be set such that it may be available publically, allowing other users to read an individual’s comments on a document. This has numerous applications for users of all levels of expertise. For instance, experienced users may choose to create public notes and annotations which can expand on the text of the document, helping less experienced users to comprehend the content. Less experienced users may indicate parts of the document which they would like to be explained further. In this way, CAT can play a crucial role in collaboration, especially as individual groups of users can be created, which provides a much finer granularity in between public and private annotations.

While CAT is beneficial for researchers and educators, it is also being used as an important source of user metadata for CULTURA. For instance, annotations provide an insight into which entities are of interest to a user. If a user is frequently annotating a document, it is likely that this document is of interest to them. Furthermore, by analysing the text being annotated (as well as the text of the annotation itself) using entity extraction; it is possible to discern specific items of interest to the user. Another data that the annotation tool allows to be collected is the interaction of users within a group, and how often a person views annotations by a specific contributor. This data can improve the recommendations given by CULTURA, as well as help foster collaboration within the environment. Within CULTURA, all this additional data is sent to the user model stored in MongoDB; thus, significantly increasing the metadata richness of the CULTURA environment.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.png}
\caption{Screenshot of CAT within CULTURA}
\end{figure}
6 Evaluation

Initial evaluations (Bailey et al., 2012; Ponchia, 2013) of the CULTURA environment have been performed with a number of different cohorts including professional historians, postgraduate students and members of the public. In order to assess the suitability of the CULTURA environment for users with little knowledge of the cultural collection under examination (in this case the 1641 Depositions), an important evaluation of the prototype system took place with students aged 15–17. Before the experiment started, the users were given a five-minute overview of the environment and the relevant features that they needed to understand.

The students were then divided into three teams and asked to pursue different group tasks on the Depositions which were developed by a professional historian. A typical example of the type of task set is as follows: ‘Have a look at the deposition of John Crooke and Richard Sergier who owned a bookshop in Dublin. What does their deposition tell us about life in Ireland before the rebellion? What does their deposition tell us about the book trade in Ireland? Mark up your thoughts using the annotation tool. Students used the ‘narrative’ module (discussed in Section 5.4) to help guide them through the Depositions, and answered the questions by creating individual annotations that were shared amongst their group.

After performing tasks for approximately 45 min, one member of each team used the CULTURA system and the group annotations generated during the experiment to present their findings to the room. Afterwards, all participants were asked to fill in a short user feedback questionnaire collecting qualitative feedback (open comments) on the overall, general impression of the system and a few questions to capture aspects of the CULTURA evaluation model in a general manner. Table 1 lists the questions asked and details the evaluation quality that each question was trying to measure.

As an initial question, a Smileyometer was used as a general measure of user experience/satisfaction. This Smileyometer has been proposed and utilised as a measure for fun and usability suitable for children (Sim et al., 2006; Read et al., 2002) and has been considered as a suitable ‘icebreaker’ for the questionnaire for adolescents in this evaluation study. Furthermore, aspects of user acceptance, i.e. ease of use, usefulness and behaviour intention, are collected by three individual questions allowing a general assessment of the factors of user acceptance according to the technology acceptance model as well as an overall acceptance measure (averaging over the three subscales/items). In addition, one item has been presented addressing adaptation quality.

In total, 22 students provided their feedback by completing the evaluation questionnaire. For the closed questions covered in the survey, average scores were calculated. With the exception of the user experience/satisfaction question (with a score range of 1–5) the possible score range for all questions was 1–7, with higher values indicating better results. In each case, a score above the centre point of the scale is considered a good evaluation.

An overview on the average scores is given in Table 2. As can be seen, the assessments given by the users are consistently very good. The minimum score of 1 has not been given at all for any of the questions. The mean score on the general assessment of user experience/satisfaction was 3.82 (SD = 0.73, Median = 4.0) with 5 being the maximum score available. When having a closer look at the response frequencies for the individual response options in Table 3, it can be seen that nearly 60% of the students (i.e. 13 persons) have assessed their experience as ‘really good’ and about 14% have indicated to have had a ‘brilliant’ experience. There is only one person who defined his/her experience as ‘not very good’. These results are encouraging and the overall assessment of the system was positive indicated by school students.

Table 1 Evaluation questionnaire features (see online version for colours)

<table>
<thead>
<tr>
<th>Evaluation quality</th>
<th>Question</th>
<th>Answer format/options</th>
</tr>
</thead>
<tbody>
<tr>
<td>User experience/satisfaction</td>
<td>How was your experience with the CULTURA system?</td>
<td>Free text</td>
</tr>
<tr>
<td>Qualitative feedback— overall impression</td>
<td>What did you like best about the CULTURA system?</td>
<td>Free text</td>
</tr>
<tr>
<td>Qualitative feedback – issues for further improvement/development</td>
<td>What should be improved and how?</td>
<td>Free text</td>
</tr>
<tr>
<td>Adaptation quality</td>
<td>I feel that the CULTURA system tried to support me in exploring the digital collection.</td>
<td>7-point rating scale from ‘strongly disagree’ to ‘strongly agree’</td>
</tr>
<tr>
<td>User acceptance – perceived ease of use (also usability)</td>
<td>Learning to use the CULTURA system is easy for me.</td>
<td>7-point rating scale from ‘strongly disagree’ to ‘strongly agree’</td>
</tr>
<tr>
<td>User Acceptance – perceived usefulness</td>
<td>I would find the CULTURA system useful in my school tasks.</td>
<td>7-point rating scale from ‘strongly disagree’ to ‘strongly agree’</td>
</tr>
<tr>
<td>User acceptance – behaviour intention</td>
<td>I would like to use the CULTURA system in the future.</td>
<td>7-point rating scale from ‘strongly disagree’ to ‘strongly agree’</td>
</tr>
<tr>
<td>User acceptance – behaviour intention</td>
<td>I would recommend the CULTURA system to my peers/friends.</td>
<td>7-point rating scale from ‘strongly disagree’ to ‘strongly agree’</td>
</tr>
</tbody>
</table>
In order to gain a deeper understanding of users’ personal opinions and experience of the CULTURA system, and to identify issues and suggestions for further improvement, the responses provided by students to the three open questions were analysed. The annotation tool was clearly the most popular response, when asked for the thing(s) liked best about the CULTURA system. Eleven people pointed to the possibility of making annotations, and also 11 people explicitly mentioned the possibility of sharing annotations. In particular, in this context the support that this functionality brings for collaboration was appreciated, e.g. “the way that you could annotate and also see other people’s annotations on highlighted text... this allowed us to refer and consider other people's ideas and come to a more rounded conclusion as a group”. The easy and online accessibility of resources was mentioned by four persons. Moreover, four students acknowledged the possibility of seeing the primary resources and found this very interesting, e.g. “Extremely impressed with the level of authenticity. Although the ‘old’ language was difficult to ‘de-code’ it was more interesting in its raw form”. Two people mentioned as positive that the content was normalised for searching the archive. Table 4 gives an overview over all responses and answer frequencies.

From the things liked least, the top answer was the difficulty to read and understand the deposition texts. This is actually an aspect that is completely due to the digital collection and type of cultural heritage documents covered. One student, for instance, stated that “The language in primary sources is very challenging: words are spelled differently and they have a lack of full stops that makes it hard to follow”. Four people mentioned difficulties in distinguishing annotations made by different persons or annotations that are overlapping, for example “Annotation system, although useful, became a little confusing when a room of 21 people’s annotations could all be seen at once”. Three persons criticised the inability to access the normalised resources, which is closely related to the mentioned problems in understanding and reading the early modern English. Displaying the normalised text of the resources was difficult to ‘de-code’ it was more interesting in its raw form. Two people mentioned as positive that the content was normalised for searching the archive. Table 4 gives an overview over all responses and answer frequencies.

From the other items, the best results can be identified for adaptation quality and for ease of use (usability), with average scores above 6 (7 being the maximum score available). In both cases, individual assessments of students have a score of at least 4; thus, students scored these aspects consistently very good. The results of the evaluation indicate that the CULTURA system is perceived as supporting students in their exploration of the digital collection (i.e. very good adaptation quality), and is easy to learn and use (i.e. ease of use/usability are very good). For perceived usefulness and behaviour intention slightly lower scores of 5.86 and 5.71 were calculated, respectively, which is nevertheless a very good result. Overall, students considered the system is useful in their school tasks and noted that they intend to use the system again in the future, which argues for good user acceptance.

In addition to the statistics on average scores, correlations were calculated for the questionnaire item scores. User experience correlated significantly and positively with perceived usefulness \((r = 0.802, p < 0.01)\), as well as on a medium to high level with behaviour intention \((r = 0.682, p < 0.01)\). Higher assessments of user experience were also correlated with higher assessments on perceived usefulness and behaviour intention. Furthermore, between perceived usefulness and behaviour intention a high positive correlation was identified \((r = 0.846, p < 0.01)\). Between the two items addressing behaviour intention a very high positive correlation \((r = 0.846, p < 0.01)\) was found, as would be expected. However, neither adaptation quality nor ease of use was found to be significantly correlated with any other aspect/item.

Table 2  Descriptive statistics for user questionnaire

<table>
<thead>
<tr>
<th></th>
<th>User experience/ satisfaction</th>
<th>Adaptation quality</th>
<th>Ease of use/ usability</th>
<th>Perceived usefulness</th>
<th>Behaviour intention to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.82</td>
<td>6.09</td>
<td>6.14</td>
<td>5.86</td>
<td>5.70</td>
</tr>
<tr>
<td>Median</td>
<td>4.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>SD</td>
<td>0.733</td>
<td>0.750</td>
<td>0.889</td>
<td>1.246</td>
<td>1.221</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Maximum</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3  Response frequencies for question on user experience

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not very good</td>
<td>4.5</td>
</tr>
<tr>
<td>Good</td>
<td>22.7</td>
</tr>
<tr>
<td>Really good</td>
<td>59.1</td>
</tr>
<tr>
<td>Brilliant</td>
<td>13.6</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 4  Overview of aspects liked best about the CULTURA system

<table>
<thead>
<tr>
<th>Answer category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possibility to make annotations</td>
<td>11</td>
</tr>
<tr>
<td>Shared annotations</td>
<td>11</td>
</tr>
<tr>
<td>Easy accessibility of resources</td>
<td>4</td>
</tr>
<tr>
<td>Possibility to see primary resources</td>
<td>4</td>
</tr>
<tr>
<td>Normalisation of contents</td>
<td>2</td>
</tr>
</tbody>
</table>
Depositions. The inability to view two or more resources at a time was mentioned as a negative aspect by two students, e.g. “It was difficult to compare several sources: you could only see one at a time”. Besides that two students reported difficulties in navigating through the resources and two indicated that the system was rather slow at times. Table 5 gives an overview over all responses and their answer frequencies.

<table>
<thead>
<tr>
<th>Answer category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty to read and understand the contents</td>
<td>6</td>
</tr>
<tr>
<td>Annotations are difficult to distinguish (for larger groups and/or when overlapping)</td>
<td>4</td>
</tr>
<tr>
<td>Inability to view/access normalised content</td>
<td>3</td>
</tr>
<tr>
<td>Inability to view several resources at a time</td>
<td>2</td>
</tr>
<tr>
<td>Difficulty in navigating back and forth</td>
<td>2</td>
</tr>
<tr>
<td>System is occasionally slow</td>
<td>2</td>
</tr>
<tr>
<td>No possibility/area to summarise a whole resource</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5 Overview of aspects liked least about the CULTURA system

Finally, issues and suggestions for further improvements were also asked of the school students. The top answer on this was the suggestion to provide access to the normalised resources (e.g. ‘Ability to view the sources in modern English’) or, respectively, to a built in normalisation tool, which was given by 13 persons. Five students suggested incorporating categorisation or colouring coding of annotations, such to better distinguish between them. The possibility of viewing several sources at a time was indicated as desirable by three people; and two suggested improving the GUI (e.g. ‘Aesthetic ‘fleshing out’ of website’). A range of individual suggestions were made. Partly, these refer specifically to the annotation functionality, e.g. suggesting the possibility to reply to or evaluate others’ annotations. Other comments refer to the system on a more general level, like the possibility to see the transcript and image of a deposition all at once.

Overall, the CULTURA system was perceived very positively by the evaluation participants in both, quantitative and qualitative feedback gathered. The annotation functionality and its use for collaboration were particularly appreciated, while nevertheless some weaknesses of this feature have also been indicated (like difficulties in distinguishing annotations). The suggestions for improvements collected from students provide useful information for further improving the annotation tool, as well as other features of the CULTURA environment.

7 Summary and future work

This paper discussed the CULTURA project and outlined key challenges that it is addressing within the digital humanities field. Two contrasting cultural collections (the 1641 Depositions and the IPSA collection) that have been incorporated into CULTURA were described, along with details of the service-oriented architecture underpinning the environment. Five specific features of the CULTURA environment were discussed in further detail (personalisation and adaption techniques, annotation, normalisation, entity-oriented search and network analysis), with an emphasis on the importance that metadata plays in facilitating such functionality.

A detailed evaluation with 22 students was described, which indicated the benefits that CULTURA can bring to novice users of digital cultural collections. Further evaluation studies within the CULTURA project are currently taking place, involving both the IPSA collection and the 1641 Depositions. Moreover, all the key stakeholders in this domain (from professional researchers to members of the public) will be involved. The outcomes of these studies will help refine the implementation and underlying methodology, especially in how the various models and metadata interact. Finally, the new technologies which have been developed for CULTURA (text normalisation service, SNA and INA services, entity detection service, etc.) will be fully integrated into the architecture. This will result in an end-to-end system, which encompasses all stages from the initial normalisation of cultural heritage collections, to the deployment of these resources within an online personalised portal.

References


Notes

1 http://www.bhl-europe.eu/
2 http://www.digital-collections.de/
3 http://www.europeana.eu/
4 http://dm2e.eu/
5 http://e-culture.multimedian.nl/
6 http://www.chip-project.org/
7 http://www.mosaica-project.eu
8 http://www.qviz.eu/
9 http://www.europeana.eu/portal/
10 http://www.paths-project.eu/
11 http://www.natural-europe.eu/
12 http://www.itunes.com/
13 http://www.myspace.com
14 http://www.icloud.com
15 http://1641.tcd.ie/
16 http://www.ipsa-project.org/