The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections

Cormac Hampson[1], Maristella Agosti[2], Nicola Orio[2], Eoin Bailey[1], Seamus Lawless[1], Owen Conlan[1] and Vincent Wade[1]

[1] Knowledge and Data Engineering Group, Trinity College Dublin, Ireland
{cormac.hampson, eoin.bailey, seamus.lawless, owen.conlan, vincent.wade}@scss.tcd.ie
[2] Department of Information Engineering, University of Padua, Italy
{agosti, orio}@dei.unipd.it

Abstract. In recent years there has been a marked uptake in the digitisation of cultural heritage collections. Though this has enabled more sources to be made available to experts and the wider public, curators still struggle to instigate and enhance engagement with cultural archives. This is largely due to the monolithic nature of many digital archives; the challenge of understanding large collections, especially if the language is inconsistent; and because users vary in expertise and have different tasks and goals that they are trying to accomplish. This paper describes CULTURA, an FP7 funded project that is addressing these specific issues. The various technologies and approaches being used by CULTURA are discussed, along with the lessons learnt thus far, and the future work necessary to be implemented before the project concludes.

Keywords: CULTURA, Digital Humanities, Personalisation, Adaptation, Social Network Analysis, Influence Network Analysis, Text Normalisation.

1 Introduction

The interdisciplinary field of Digital Humanities is concerned with the intersection of information communication technologies such as digital libraries, and a wide range of humanities disciplines, including history and art history [1]. Despite fresh impetus experienced in the field in recent years, current research practices in the humanities still tend to be very labour intensive, solitary and characterised by research material which is often disconnected and non-digitised. This presents a particular obstacle to both novice researchers and the general public. Widespread digitisation represents an important step forward. However there remains a real need for specialist environments which offer a rich, personalised and stimulating engagement with this digitised material. A key challenge still facing curators and providers of digital cultural heritage is to instigate and enhance engagement with digital libraries that manage cultural collections. CULTURA [2] is a three year, FP7 funded project scheduled to finish in February 2014. Its main objective is to pioneer the development of personalised information retrieval and presentation, contextual adaptivity, and social analysis in a digital humanities context.

Personalisation and adaptive contextualisation technologies such as adaptive hypermedia, adaptive web, intelligent systems, personalised information retrieval systems and recommendation systems have been successful in many application areas such as education, tourism, and general information sites [3]. These technologies reconcile each user's interests, prior experience or location to provide personalised navigations of relevant digital resources (adaptive personalisation) or to suggest personalised recommendations based on similar users' behaviour and feedback (social recommendation). However, current adaptive personalisation technologies typically have two key weaknesses: They fail to take into account any broader community of which the user is a member, thus neglecting a valuable source of insight into user intention; They are unaware of the structure and internal dynamics of the material to which they offer access. Such "domain awareness" is an important input to the selection and sequencing of material presented by an adaptive system to the user.

Section 2 describes the CULTURA project and how the various partners are contributing to enhanced user engagement with cultural archives. Section 3 discusses the technological improvements that CULTURA is developing, with section 4 highlighting the key lessons learnt thus far in the project. Finally section 5 summarises the paper and briefly discusses the future work that will be undertaken.

2 The CULTURA Project

The CULTURA project brings together complementary expertise, which is required to address the different goals described in the previous section. In particular, there are three main areas to which CULTURA partners contribute: (i) reference collections related to cultural heritage; (ii) innovative content processing techniques; and (iii) evaluation/validation of the results through user studies. Two cultural heritage collections are the focus of case studies: the 1641 Depositions [4], provided by scholars of the Department of History at the Trinity College Dublin; and the Imaginum Patavinae Scientiae Archivum (IPSA) [5] digital archive of illuminated manuscripts, provided by scholars in history of art of the Department of Cultural Heritage and researchers in computer science of the Department of Information Engineering, both at the University of Padua.

Both collections are the subject of scientific research by communities of experts, and represent two complementary approaches to cultural heritage research. On the one hand, the 1641 Depositions allow historians to study original witness testimonies and the network of relationships between rebels, deponents, locations, and timings. On the other hand, the IPSA digital archive allows are historians to study artifacts, and the network of relationships between authors, illuminators, and visual resemblances. A key factor of the CULTURA project is the direct involvement of both groups of humanities scholars, who are active members of the CULTURA consortium.

With the aim of providing adaptive and personalized access to the two collections, the CULTURA project addresses a number of relevant research issues. First of all, historical texts need to undergo a normalization process, which is undertaken by researchers of Sofia University, before exploiting automatic analysis techniques. For

instance, the names of persons cited in the 1641 Deposition or the descriptions of plants drawn in illuminated manuscripts of IPSA, are reported with many variants that need to be normalized. Automatic text analysis is mainly carried out IBM Israel.

Special attention is given to exploiting the network of relationships that are typical of cultural heritage collections. In particular, techniques of social network analysis and influencer network analysis are developed by another industrial partner, Commetric. These techniques allow personalized and adaptive visualization of cultural heritage collections that, combined with social search techniques developed by IBM, form a core component of the CULTURA portal. The development of the portal, which will grant adaptive and personalized access to both collections, is designed and implemented by the coordinator of the project: the Knowledge and Data Engineering Group of Trinity College Dublin. An overview of the technical aspects of the CULTURA approach is described in the next section.

The cohort, or role of the user, is a crucial aspect that must be considered when developing cultural heritage applications. Both the 1641 Depositions and IPSA are of interest for a variety of user cohorts, ranging from the domain experts (who carry out scientific research and produce original results on the collections) to scholars in other domains, students, and the general public (who have a variable interested in accessing to cultural heritage content). Hence, a large amount of effort is focussed on the development of user studies and evaluations, which are coordinated by the Knowledge Management Institute of the Technical University of Graz. These user studies involve active participation from scholars of the University of Padova and Trinity College Dublin.

3 The CULTURA Approach

Central to the CULTURA project is the development of a corpus agnostic environment that enables different cultural collections to exploit the various tools and services it offers. This approach means that any cultural archive integrated with CULTURA benefits from an interactive user environment which dynamically tailors the investigation, comprehension and enrichment of the collection to the individual user. In order for CULTURA to support such services, its research is focussed on advancing and integrating several key technologies including:

- cutting edge natural language processing, which normalises ambiguities in noisy historical texts.
- entity and relationship extraction, which highlights the key individuals, events, dates and other entities and relationships within unstructured text.
- social network and influencer network analysis of the entities and relationships within the content, and also of the individuals and broader community of users engaging with the content.
- multi-model adaptivity to support dynamic reconciliation of multiple dimensions of personalisation.

Through the provision of such services, CULTURA empowers users with different levels of expertise to investigate, comprehend and contribute to digital cultural

collections. How each of these technologies is being progressed in CULTURA is now discussed in turn with reference to one the cultural collections, the 1641 depositions, which is the initial cultural archive being integrated with CULTURA. The 1641 depositions are an unparalleled source anywhere in early modern Europe. They consist of over 8,000 witness testimonies, from all social backgrounds, concerning their experiences of the 1641 Irish rebellion. Typically they document the loss of goods, military activity, and the alleged crimes committed by the Irish insurgents.

3.1 Text Normalisation

A significant problem in many cultural collections, such as the 1641 depositions, is that they contain noisy text with inconsistent references to, and spellings of, the same entity e.g. person, place, event. The depositions are an extreme example of this, with the particular form of early modern English used displaying massive inconsistencies across the collection. CULTURA is addressing this with an initial normalisation of the collection before it enters the workflow. The primary purpose of normalisation is to produce documents without historical variation. The elimination of such variations results in documents that are more easily processed and analysed than the originals. At this point it should be clarified that the unnormalised text of the collection is still available to scholars who use the CULTURA portal and that every effort is made during normalisation to ensure that changes that implicitly involve interpretation of the text, or that go beyond normalisation, are avoided. In terms of text normalisation within the CULTURA project, a number of important advancements have occurred including the development of:

- a new normalisation algorithm [6] called Regularities Based Embedding of Language Structures (REBELS).
- an integrated REST based web service with user friendly interface for the implemented normalisation module
- a tool for manual annotation which makes the manual normalisation process as simple as possible, and helps to verify consistency of the annotations and to help the resolution of detected conflicts.

These developments all support a more effective text normalisation process and improve the effectiveness of the entity and relationship extraction procedures which are described next.

3.2 Entity and Relationship Extraction

In most cases, entities are at the core of network analysis studies and the focus is on identifying the most, and least, important entities in a network. An entity can be anything that interacts with or has a relationship with another entity. In terms of the 1641 Depositions, entities include people, events and locations mentioned within the text. The automatic extraction and normalisation of entities has been enhanced by the

advancement of Natural Language Processing (NLP). This is often helpful in processing large volumes of data very quickly, but 100% accuracy cannot be guaranteed. Additionally, the development of NLP engines and libraries for specific entity types often require extensive human effort and a lot of time.

Once the entities have been mapped onto a network, the links between nodes indicate a relationship between two entities. This is done using the SaND tool [7] developed by IBM Israel. For the initial deployment of the CULTURA portal these entities and links are from the 1641 Depositions. From a single text several relationships between the mentioned entities can often be identified, this provides for several different networks being produced and studied. For example, in the 1641 Depositions, the people mentioned in it can be linked using different criteria such as: living in the same town, having the same occupation, committing a similar crime etc. Once entities are extracted from the cultural collection, and links between entities are identified, social and influence analysis can be applied, which is a key feature of CULTURA and is described in the following section.

3.3 Social Network Analysis (SNA) and Influencer Network Analysis (INA)

Social Network Analysis (SNA) and Influencer Network Analysis (INA) are being conducted within CULTURA to identify the communities and relative influence of individuals described in the content collections. Taking the 1641 depositions, these enhanced, annotated collections enable the implementation of features such as:

- Similar Entity Recommendation, e.g. Event A similar to Event B as both are murders which occurred in a particular place at a particular time
- Parent Entity Recommendation, e.g. County A recommended as it contains region B)
- Same-Entity Different-Label recommendations, all references to a particular person resolved across documents, including resolution of spelling variation, use of patronymics etc.
- Related Entity Recommendation, e.g. same participants, sequences of events

Furthermore, the SNA and INA are not only applied to the content collections as described above, but also on the communities of researchers surrounding the collections. The community-based adaptive service will use the output of this analysis to offer features such as:

- discovery of relevant experts to connect users
- aggregation of tasks around particular communities
- creation of new communities based on similar activities
- support for community collaboration and contribution
- community-created narratives (guides)

The provision of SNA and INA offers users of the CULTURA portal new methods to interrogate the underlying cultural collection, as well as visualising their community

of peers and identifying potential avenues for collaboration. One aspect that is vital in CULTURA is that appropriate recommendations and supports are provided for users of different expertise which is one of the challenges addressed by the multi-model adaptivity methodology which is described in the next section.

3.4 Multi-model Adaptivity

The CULTURA portal, through which end users interact with the underlying cultural collection, is a next generation adaptive system that provides multi-dimensional adaptivity along several "axes". These axes include:

- a personalised information retrieval and presentation system which responds to models of user and contextual intent.
- community-aware adaptivity which responds to wider community activity, interest, contribution and experience.
- content-aware adaptivity which responds to the entities and relationships automatically identified within the artefacts and across collections.
- personalised dynamic storylines which are generated across individuals as well as entire collections of artefacts.

The Adaptive Engine is an independent service from the CULTURA portal, which is responsible for reconciling these various axes of personalisation and providing relevant information to the user that will enhance their experience within the portal. To achieve this, a model of each user is built silently as a user interacts with the system. A user will have the opportunity to scrutinise, edit and enhance their own personal user model if they feel it does not accurately represent their interests. All actions a user performs are recorded in order to provide detailed information on each user. In terms of the 1641 Depositions, this includes viewing a deposition, bookmarking a deposition, and annotating a deposition. Other user information will also make up the model, such as the experience a user has with a collection, and the communities they are a part of. This detailed model of a user helps CULTURA ensure that the recommended information and services are both broadly and contextually relevant for the user.

4 Lessons Learned

One of the main guidelines in the development of the CULTURA project is to consider that multimedia resources for cultural heritage play a dual role. On the one hand they provide an environment for scholars to develop novel research relating to the collections, as well as keeping track of their findings within the same digital environment. On the other hand, they play a major role in the dissemination of cultural heritage to a wider audience who might have only a marginal interest in the content of the collections, and are looking to improve their knowledge. The need to

support both these approaches to interacting with cultural heritage collections was clearly evident from the initial evaluation [8] and user requirements collection.

Expert users already have a deep knowledge of the content of the digital collections allowing their research to create new insights into these artefacts. Thus, historians studying documents of the 1641 Depositions and art historians studying illuminated manuscripts of the IPSA digital archive are well aware of the relevant search parameters to access the digital content and are motivated to interact with the system because of their research interests. In contrast, other user groups that participated in the evaluation (mostly students and scholars in related disciplines) needed to be guided, through user-friendly interfaces, in order to reach relevant digital content. Difficulties in accessing digital content easily can result in a loss of interest towards the collection in question. Fortunately, the initial evaluation of the CULTURA environment did highlight an important direction for improving the experience of non-domain users with cultural heritage collections. For instance, most of users who were interviewed described as particularly relevant the role of scholars as mediators between the general public and the digital content. The fact that both collections were actively studied by professional researchers was considered an important factor by non-domain users, who were interested in understanding how the research process was carried out, and to retrieve additional information to give them a better understanding of the digital content.

The role of scholars as mediators between cultural heritage collection and the general public was somewhat predictable. However, what was perhaps less obvious was the importance, for non-domain users, of accessing the same collections and using the same tools that the professional researchers use. For instance, although images of illuminated manuscripts are beautiful to see and historical depositions are very interesting to read, a sustained interest is more likely when non-domain users are facilitated in reaching a deeper understanding of the digital content, thanks to the work of scholars. Thus a key aspect of the CULTURA project is to improve adaptivity and personalisation techniques so that the same environment can be used effectively by users with very different levels of expertise and motivation.

Another lesson that can be learned from the initial user study regards the difficulties in providing a research environment that can be useful for scholars in different disciplines. A substantial amount of collaborative work carried out in the initial phase of the project was devoted to the harmonization of research needs of historians and art historians. Although the two groups of scholars are interested in different kinds of documents, there are a number of common characteristics that will be taken into account in the next phases of the project. In particular, the need to analyse the relationships between elements of the digital collection e.g. persons, artists, locations, or subjects of drawings, is shared by both groups of scholars. Thus, the integration of automatic tools for social network analysis and influencer network analysis, which is already one of the main goals of the CULTURA project, becomes a crucial point for success of the project, and will be addressed in phase two and three of the project. The integration of two different collections in the same environment provided by CULTURA will be also enable a number of best practices to be described, so that project results can be reused, and other collections (and involving an increasing number of cultural heritage custodians) can be easily integrated.

5 Summary and Future Work

This paper has described the CULTURA project and how it is addressing the need for next generation digital research environments to support engagement with cultural collections. Phase one of the CULTURA project has already been completed and an initial environment deployed containing the 1641 Depositions. Integration of a second collection, Imaginum Patavinae Scientiae Archivum (IPSA) into CULTURA, is in progress. From a technical perspective, IPSA represents a very different kind of cultural archive to the 1641 Collection. The IPSA collection is primarily image rather than text based, and has 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.

The next stage of evaluation in CULTURA will involve both the IPSA collection as well as the 1641 Depositions. Furthermore, all the key stakeholders in this domain (Professional researchers, amateur researchers, history enthusiasts and members of the public) will be involved, and the outcomes of these studies will help refine the implementation and underlying methodology for the next phase of the CULTURA project. Finally, in the next phase of the project, the new technologies which have been developed within CULTURA (text normalisation service, SNA and INA services etc.) will be integrated into an overall service-orientated architecture. This will enable an overall workflow, with seamless service integration, which goes from the normalisation of the cultural collection to the deployment of these artefacts within a user friendly adaptive portal.

Acknowledgments. The work reported has been funded by the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement no. 269973.

References

1. Agosti, M., and Orio, N.: The CULTURA Project: CULTivating Understanding and Research through Adaptivity. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) Digital Libraries and Archives, vol. 249, pp. 111--114. Springer, Heidelberg (2011)
2. The CULTURA Project, http://www.cultura-strep.eu
3. Agosti, M., Lawless, S.: The CULTURA project: CULTivating Understanding and Research through Adaptivity. In the Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval, PMHR 2011, Eindhoven, Netherlands, pp.50—54,(2011)
4. 1641 Depositions, http://1641.tcd.ie
5. IPSA Imaginum Patavinae Scientiae Archivum, http://www.ipsa-project.org
6. Gerdjikov, S.: Some algebraic properties of alignments of words. In Comptes rendus de l'académie bulgare des science. (in press 2012).
7. SaND Project, http://www.research.ibm.com/haifa/projects/imt/social/sand.shtml
8. Bailey, E., Lawless, S., O'Connor, A., Sweetnam, S., Conlan, O., Hampson, C., Wade, V.: CULTURA: Supporting Enhanced Exploration of Cultural Archives through Personalisation. In the Proceedings of the 2nd International Conference on Humanities, Society and Culture, ICHSC 2012, Hong Kong, China. (in press, 2012).