CULTURA: Supporting Professional Humanities Researchers

Eoin Bailey, Mark Sweetnam, Michael O'Siochru, Owen Conlan

A key challenge facing professional researchers in the domain of cultural heritage across Europe and worldwide is the interrogation of growing digital humanities collections. 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 highly variable in terms of language, spelling, punctuation, and consistency of terminology. These difficulties are compounded by a lack of normalised spelling in most European languages before the eighteenth century. This means that search across these digital collections tends to return sub-par results as multiple spellings for many common words are treated as independent document keywords. CULTURA is a corpus agnostic environment with a suite of services, including personalisation, annotation, and recommendation, providing necessary supports and features for a diverse range of professional researchers.

In order to 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 is required to be integrated and intelligent. Such next generation adaptivity, as espoused by CULTURA, supports the dynamic composition and presentation of digital cultural heritage resources. Automated adaptivity however, is not enough on its own. Ensuring that the user is in control of the personalisation process is essential. Such user-centred control is enhanced through correlating usage patterns with self-expressed user goals; pre-defined strategies (e.g. research strategies, investigation strategies, discovery strategies, explanatory strategies etc.); and the provision of appropriate tools for users to explore and navigate large cultural heritage information spaces.

A central aspect of the CULTURA environment is its use of rich metadata (user generated, computer generated, and expert generated) coupled with natural language processing, entity extraction and social network analysis techniques, in order to support collaborative exploration, interrogation and interpretation of the underlying cultural resources. The Qviz [1] 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, however Qviz, however, does not incorporate a personalised or adaptive aspect.

The manual determination of descriptive metadata across a large corpus is too time-consuming to be practical. For example, the process of metadata identification for the 8,000 1641 depositions took a research assistant 12 months [2]. An automated entity extraction process is used by the CULTURA project [3]. This process interprets words and combinations of words to identify entities in the corpus such as people, places, events, and dates. Complex entities are then constructed from these entities and the corpus allowing the identification of events, such as *WHO* did *WHAT* on such a *DATE* at a specific *LOCATION*.

Entity extraction is most powerful on a corpus in which all the content has been normalised as entities can be matched across multiple documents. To enable normalisation a ground-truth is manually generated across a proportion of the corpus, approximately 10% in the case of the 1641 depositions [4]. The ground-truth assigns non-normalised terms in the corpus to normalised terms. The ground-truth is applied in the generation of a statistical model that is then utilised on the entire corpus outputting a normalised corpus. It is this normalised content from which entities are identified and extracted. The linguistic model used in the normalisation of the 1641 Depositions has proved highly reusable, and provides robust results when applied to a range of other material contemporary to the corpus used to generate the model, enabling the re-use of the model on additional content collections.

The output of these processes ensures each individual piece of content from the corpus has descriptive metadata in the form of individual entities such as a person or place, and complex entities such as where a person lived i.e. compound entities, as well as a normalised variant. This enables a simple keyword search to provide results that cover the entire corpus and all variants of spelling of the terms entered. Additionally, and more significantly, individual pieces of content from within the corpus can be linked to other content that contains information on the same event or similar events, based on the date, location, people, and type of event. These links enable professional researchers to quickly identify content that is related to their current research topic, by way of visualisations of these links [4].

While the automated tools are providing useful results for professional researchers, these state-of-the-art tools cannot replace the insight and experience of a professional researcher. These insights can be captured via an annotation tool that can be used to annotate specific aspects of any piece of content, both textual and visual. Annotations can be at any level, from a single word up to the entire document, or any user-identified region of an image. Annotations also allow the researcher to link the identified content to any other document within the corpus. Links such as this feed a professional researcher's knowledge into the system, and can be used to aid in the adaptivity and personalisation of the system for all users.

Collaboration between professional researchers occurs in an implicit manner via the use of annotations as link generation between content, which feeds into recommendations and personalisation. Explicit collaboration is also present in CULTURA. Researchers can share annotations with other users, enabling the propagation of insights and discoveries in the content. As annotations are anchored to specific elements of a document they provide a powerful mechanism for inserting detailed and relevant knowledge. This knowledge can be made available to either groups of researchers or all researchers thus enabling a greater collaboration across all users of the environment.

The CULTURA environment has already been used by professional researchers in the course of their research. Services including annotations, document level notes, and multi-dimensional recommendations of content were enabled within the environment.

A number of the professional researchers involved in trials of the CULTURA environment were initially wary of the potential impact of adaptivity on their research. They expressed concern that the recommendation system would create a 'filter bubble', distorting the appearance of the collection. As they engaged with the environment, however, these concerns were abated. Research into the user requirements of these end-users had identified this concern, and a concomitant need for a high level of scrutability. In line with this, recommendations made by the CULTURA engine were very explicitly presented in a way that made it easy for researchers to decide whether to use them, or to ignore their suggestions. In addition, the interface made it clear why a given user was being recommended a particular piece of content.

After some initial caution, professional researchers agreed that the adaptive recommendations provided by CULTURA were genuinely useful. This was especially true of researchers who were previously unfamiliar with the CULTURA corpora. These users reported that the recommendations facilitated their mastery of the collection. Researchers who had greater familiarity with the collection rated the utility of the recommendations less highly, but still expressed an appreciation of their potential usefulness. In particular, they suggested that the recommendations were valuable in encouraging them to look at the collection in new ways.

The user model for the professional researchers was evaluated and determined to be an accurate reflection of the researcher's interests and and topics under research. While some anomalies were identified in the user model, the areas of interest that were weighted most highly correlated with the topics of the research.

The CULTURA environment is currently engaged in additional evaluations with trainee researchers utilising a broader set of documents, tools, and adaptivity.

[1] http://www.qviz.eu

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

[3] Carmel, D., Zwerdling, N., Yogev, S., Entity oriented search and exploration for cultural heritage collections: the EU cultura project. In Proceedings of the 21st international conference companion on World Wide Web. ACM, New York, USA, 227-230 (2012).

[4] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., Wade, V., CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. EUROMED 2012, International Conference on Cultural Heritage. Lemesos, Cyprus. To Appear.