The CULTURA Portal

Exploring Cultural Treasures

Gary Munnelly, Cormac Hampson, Seamus Lawless, and Owen Conlan

Knowledge and Data Engineering Group, SCSS, Trinity College, Dublin 2, Ireland
{Gary.Munnelly, Cormac.Hampson, Seamus.Lawless,
Owen.Conlan}@scss.tcd.ie

Abstract. This paper introduces the CULTURA system which is pioneering the next generation of online tools for interacting with the cultural treasures of Europe. An overview of the architecture is presented which highlights some of the key features of the CULTURA environment. This is accompanied by a brief description of the intended workflow of both the user and the services. A live version of the portal can be found at http://cultura-project.eu

Keywords: CULTURA, personalization, digital heritage, user modelling.

1 Introduction

The information age has provided humanity with a level of access to knowledge that is incomparable with any other period in recorded history. Documents about almost any subject, from detailed information about the life of Henry VIII to instructions regarding how to eat an orange can be located and consumed within moments by those who are interested in such things.

This information explosion has resulted in an empowered community of users who are confident that when they wish to learn about a subject, relevant data is little more than a Google search away. However, some resources that have remained inaccessible to the average individual include many of the original, primary sources of cultural heritage. While efforts have been made to digitize these documents, either by scanning the originals or transcribing the text, the content itself can present a barrier to the would-be scholar.

Depending on the specific era, such documents may be challenging due to the density of their information, their inconsistent use of language, archaic spelling or terminology, the assumption of a certain amount of prior knowledge on the part of the reader and more [1]. This is an unfortunate circumstance as users who wish to study these texts can locate them by traditional means of searching, but often lack the tools to consume them in a more forensic manner. This paper describes CULTURA [2], a dynamic, customizable web portal which provides a suite of tools designed with the goal of empowering and assisting the user in their exploration of these cultural treasures.

2 An Overview of CULTURA

CULTURA is a three year, FP7 funded project, whose main objective is to pioneer the development of personalized information retrieval and presentation, contextual adaptivity,and social analysis, all in a digital humanities context. To that end, it employs a wide array of tools and services which are designed to aid and inform the user in their exploration of digital collections. At present, CULTURA is being trialled using two digital cultural collections – the 1641 Depositions from Ireland and the Imaginum Patavinae Scientiae Archivum (IPSA) from Italy. These collections present very different challenges for CULTURA, due to the depositions being textual in nature and IPSA being largely pictorial.

The architecture of CULTURA is service oriented, allowing the portal to be tailored to suit a particular collection. For example, if inconsistent language is not a problem in a corpus of documents, then normalization may not be a required component and can be decoupled from the rest of the site as required.

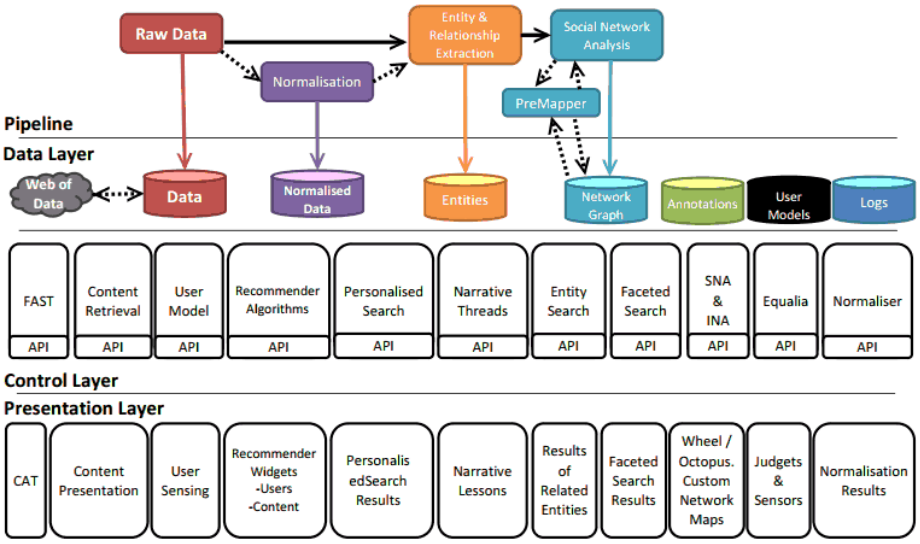

Fig. 1. Overview of the CULTURA Architecture

2.1 Pipeline and Data Layer

Before the data is exposed to the user, it undergoes some preprocessing to extract meaningful information which can be used by both the system and the researcher.

The raw data, (which is comprised of historical documents, text etc.) is passed in both its original and normalized form to the entity extraction service. The process of normalization is intended to introduce consistency into the document language by resolving some of the variant spelling into a more modern form. This is particularly important for textual corpora in which the language is so archaic that standards are

absent even within the work of an individual scribe. However, as a result of this process, some information can be lost from the artefacts, (e.g. Smythe being normalized to Smith, despite being a distinct and different name), hence the parallel analysis of the text in its original form.

Entity extraction is employed to identify the named entities within the texts including people, places, dates, etc. These entities provide an important insight into the nature of the text and can be used to guide a user towards resources which are relevant to their research as well as link documents which cover similar or related subjects. Social network analysis (SNA) is also applied to the output of this process in order to help amalgamate multiple individual entities into a single overall reference. For example, it may be possible to discern that the Phelim O'Neil whose activities in Louth are extensively documented in 1641, is the same man whose execution in 1653 is recorded elsewhere. The PreMapper tool developed in CULTURA provides a means by which the data curator can manually establish links which were not automatically detected by CULTURA.

As can be seen in Figure 1, the data layer is largely a repository for the information extracted by the pre-processing layer. Data such as the normalized version of the corpus, entities identified and how these entities relate to one another (SNA) are stored along with the original copies of the source material. This data is drawn upon by the remaining layers of the architecture for their respective purposes.

The importance of the underlying user model should also be noted. This important feature of CULTURA is a major driving force behind many of the components in the control layer. Information about a user's browsing history within the site, inferred research interests and exhibited level of expertise is persistently stored in the data layer, thus allowing a user's personalized experience to span several sessions.

2.2 Control Layer and Presentation Layer

The control layer services run as components of the live site and are interfaced with via a series of API calls from the client's computer. They can be invoked by the user to produce an effect within the CULTURA portal, e.g. annotation, normalization, etc. or they may be availed of by CULTURA itself as it attempts to personalize the user's experience, e.g. content retrieval, content recommendation, user model modification etc.

The user model is CULTURA's representation of the user. It is constructed based on a user's behaviour within the site in an attempt to discern how familiar they are with the source material, what entities and documents within a collection interest them etc. Based on this representation, CULTURA can tailor the experience of the individual to suit, not only their research interests, but also their level of expertise. For example, for a user who is exhibiting a particular interest in County Louth, both by bookmarking documents which relate to it and annotating bodies of text which contain references to it, CULTURA will attempt to establish what aspects of Louth are of interest to them by correlating the user model with the entities extracted by the pipeline. The relationships determined by entity extraction can then be used to produce

lists of alternative sources which may interest the user. These are presented to the user in a recommender block and also influence the results of personalized searches.

The presentation layer is the user facing façade of the portal though which they are given transparent access to the various services that CULTURA provides. Simple user interface controls allow users to interact directly with documents and the environment in general through annotating, sharing, bookmarking and searching. SNA widgets such as the Wheel and the Octopus can be used to visualize the entities within a resource and view the links which relate a particular entity to another in a separate document. Through methods such as this a researcher can explore the vast range of related entities which chain documents together.

User's seeking a more guided, tutorial based experience of a corpus can avail of the user narratives [3]. These narratives can be used to guide a user through a collection, explaining the content along the way and providing insight into the nature of the sources. The control layer contains a collection of narrative threads, designed by experts in the domain of the source material. These threads can be selected by the user in the presentation layer and are used to guide the individual through a tour of the corpus. The level of detail of this tour is dependent on the user's level of expertise as determined by the user model. For example, a novice user may require a more detailed tour due to their lack of familiarity with the source material.

3 Conclusion

Although it is still under development, to date the CULTURA portal has received very positive feedback during user trials regarding its usefulness as a research tool. Further enhancements are expected to increase usability and generally improve the user's experience with the portal. A current version of the site can be found at http://cultura-project.eu.

4 References

1. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica Annette Gotscharek Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, Klaus U. Schulz 2009 Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data
2. The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. C. Hampson, M. Agosti, N. Orio, E. Bailey, S. Lawless, O. Conlan, V. Wade Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.
3. The Narrative Approach to Personalisation, Owen Conlan, Athanasios Staikopoulos, Cormac Hampson, Seamus Lawless & Ian O'Keeffe, New Review of Hypermedia and Multimedia [In Press]