0.1 From research data repositories to virtual research environments: a case study from the Humanities

The difference in scholarly practices between the sciences and the mainstream humanities is highlighted in a study (Palmer *et al.*, 2009), which investigated the types of information source materials used in different humanities disciplines, based on results contained in the US Research Libraries Group (RLG) reports. Structured data is relatively little used, except in some areas of historical research, and data as it is traditionally understood in the sciences, i.e. the results of measurements and the lowest level of abstraction for the generation of scientific knowledge, even less so. It is true that the study is partly outdated, containing results from the early 1990s, and that data in the traditional sense is becoming increasingly important in the humanities, particularly for disciplines such as linguistics and archaeology in which scientific techniques have been widely adopted. Nevertheless, it is clear that in general humanities research relies not on measurements as a source of authority, but rather on the provenance of sources and assessment by peers, and that what data repositories are for the sciences, archives are for the humanities.

Indeed, studies of humanities scholars (Duff *et al.*, 2004) have demonstrated that they continue to rely on primary materials held in dedicated collections in special places, in repositories and archives, and it is in repositories (and archives) that the scholar carries out the work of assessing these source materials. In the UK and elsewhere there are significant digitisation programmes for humanities material, which to an increasing extent are able to provide the humanities researcher with digital surrogates for the physical archives. In some cases major memory institutions are systematically digitising the material for which they are responsible, but nevertheless digitisation is on the whole a somewhat piecemeal affair, and is carried out to different extents (e.g. image only or image plus OCR) and quality levels, depending on the availability of funds. Individual projects may address a particular set of archival material relating to a particular research topic, resulting in numerous dispersed (albeit usually online) resources, developed using different technologies and standards. Archival material is thus made easier to access, creating new possibilities for the researcher, but on the other hand this very availability raises new issues.

Our work sets out to investigate how (digital) repository content can be delivered to humanities researchers more effectively, independently of the location and implementation of that content, and with special means provided for customising the retrieval, management and manipulation of this information. Thus, our work is driven in part by our interpretation of the requirements from (Duff *et al.*, 2004), as they relate to enhanced methods of research on archives. Retrieval is to happen in near real time, and traditional finding aids are to be complemented by more sophisticated retrieval means. In particular, the personal copy of a finding aid that is often quoted as an important prerequisite for specialised research in archives is complemented by the ability to create on demand relevance indexes on the unstructured resources, and to combine the resources in new ways. We consider this to be the grand integration challenge for research repositories in the humanities, delivering data-driven humanities.

Many specialised Virtual Research Environments (VREs) (Fraser, July 2005) that integrate digital repositories with tools and services to work with the data in them have been developed to address particular tasks in various humanities disciplines. For example, the Silchester Town Life Project VRE (`http://www.silchester.rdg.ac.uk/`) and the subsequent Virtual Environments for Research in Archaeology (VERA)

(http://vera.rdg.ac.uk/) address data integration in archaeological excavations, while the VRE for the Study of Documents and Manuscripts (http://bvreh.humanities.ox.ac.uk/) developed services for sharing and annotating manuscripts. The King's College London-based TEXTvre (http://textvre.cerch.kcl.ac.uk/) project is concerned with the integration of institutional repositories and VREs in the specialised domain of Digital Humanities, specifically the creation of XML-based resources.

Building on the experiences of these VREs, we are addressing how to move beyond support for specific, focused tasks, and instead build services and environments that enable more general-purpose humanities research activities. The aim of our work is to find new ways of integrating and organising the heterogeneous and often unstructured digital resources in repositories used in humanities research, including advanced search and browse services required to support 'active reading' (Brockman *et al.*, 2001) processes, and to deliver a framework for the future on-demand delivery of VREs to various humanities research communities in Europe. This paper describes experiments carried out to this end as part of the ESFRI project DARIAH (http://www.dariah.eu), which aims to conceptualise and build a virtual bridge between humanities and arts resources across Europe. Subsequently, we received funding under the JISC Rapid Innovation programme for the recently started gMan project, which is consolidating these experiments.

Our starting point was *D4Science* (http://www.d4science.eu), a production-level infrastructure serving mainly scientific communities, but which is not biased towards any particular discipline and has great potential for meeting the needs that we have identified for building VREs by combining repositories resources. *gCube* (http://www.gcube-system.org), on which the infrastructure is based, is a distributed, service-based system designed to support the full life-cycle of modern research, with particular emphasis on application-level requirements for information and knowledge management. In gCube, VREs can be interactively designed and configured on demand (Candela *et al.*, 2009), and the system is responsible for its physical deployment and correct operation in the infrastructure. Computational resources are exploited for computationally demanding tasks such as on-demand indexing of large collections.

gCube application services offer a full platform for distributed hosting, management and retrieval of data and information, and a framework for extending state-of-the-art and on-demand indexing, selection, extraction, description, annotation and presentation of content. Each D4Science VRE, generated using gCube, makes available a grid-based repository to store, share and access information, a grid-based computing environment to efficiently run data analysis services and a reporting tool to publish and share information. gMan demonstrates how humanities data sets can be fed to GCube and will enhance the environment with further research services according to the needs of the humanities research community. gCube be easily extended, as it fully complies with web service standards (SOAP, BPEL, WSRF, WS-* and JSR168 Portal and Portlets). To avoid overhead with setting up the infrastructure, we shall use the existing gCube installation on the European grid infrastructure.

We are investigating how humanities repository resources can be imported into gCube, and how the VRE can be enhanced with further services according to the needs of the targeted research community. The gCube system is designed for extensibility; communities are encouraged to tailor the functionality to their particular needs, by developing new services or plugins. As initial test datasets for our experimental scenarios, we are using the three resources:

1. The Heidelberger Gesamtverzeichnis (HGV) der griechischen Papyrusurkunden Ägyp-tens (`http://www.rzuser.uni-heidelberg.de/~gv0/`), a database of metadata records for some 55,000 Greek papyri, mostly from Roman Egypt and its environs. The metadata includes (among other information) bibliography, keywords, dates and places (e.g findspots and provenances), as well as links to the corresponding documents in the Duke Databank of Documentary Papyri.

2. Projet Volterra (`http://www.ucl.ac.uk/history2/volterra/`), a database of Roman legal texts, and associated metadata, from various sources (epigraphic, papyrological, or literary). The database is currently in the low tens of thousands of texts, but very much in progress, and is stored in a series of themed tables in Microsoft Access;

3. The Inscriptions of Aphrodisias (InsAph) (`http://insaph.kcl.ac.uk/`), a corpus of about 2,000 ancient Greek inscriptions from the Roman city of Aphrodisias in Asia Minor, including transcribed texts and metadata marked up using EpiDoc TEI, as well as images of the physical objects.

We are supplementing these with three online resources, which are essentially collections of "things" — respectively places, personal names, and coins — each of which is identified by a stable URL that resolves to a systematic representation of the corresponding "thing":

4. The Pleiades Project (`http://pleiades.stoa.org/`) is based on the Barrington Atlas (`http://atlantides.org/batlas/`) and provides a catalogue for ancient places. Each is associated with a dedicated URL such as $http://pleiades.stoa.org/places/221986/aphrodisias/$.

5. The Lexicon of Greek Personal Names (LGPN), which exposes ancient Greek personal names as URLs that resolve to a representation of information about the name in either XML, JSON or RDF.

6. The American Numismatic Society's collection of coins, whose entries can be referenced by URLs that return HTML, or by DNIDs, e.g. $numismatics.org : 1933.23.1$.

The original three datasets were selected because of the diversity of their implementations and because, while originating from quite different research projects, there is a significant overlap in their contents, both in terms of places, time periods and people. They thus allow realistic cross-resource searches and queries. The supplementary resources provide useful domains for annotations and inter-object links, as there are numerous potential connections with the first three datasets.

Bibliography

Brockman, W., Newmann, L., Palmer, C. & Tidline, T. 2001 Scholarly work in the humanities and the evolving information environment.

Candela, L., Castelli, D. & Pagano, P. 2009 On-demand Virtual Research Environments and the Changing Roles of Librarians. *Library Hi Tech*, **27**(2), 239–251.

Duff, W., Craig, B. & Cherry, J. 2004 Historians use of archival sources: Promises and pitfalls of the digital age. *The Public Historian*, **26**(2), 7–22. (doi:10.1525/tph.2004.26.2.7)

Fraser, M. July 2005 Virtual research environments: Overview and activity. *Ariadne*, (44).

Palmer, C. L., Teffeau, L. C. & Pirmann, C. M. 2009 Scholarly information practices in the online environment: Themes from the literature and implications for library service development. report commissioned by oclc research. `www.oclc.org/programs/publications/reports/2009-02.pdf`.