

Deploying general-purpose virtual research environments for humanities research

Tobias Blanke, Leonardo Candela, Mark Hedges, Mike Priddy and Fabio Simeoni

Phil. Trans. R. Soc. A 2010 **368**, 3813-3828

doi: 10.1098/rsta.2010.0167

References

[This article cites 6 articles](#)

<http://rsta.royalsocietypublishing.org/content/368/1925/3813.full.html#ref-list-1>

Rapid response

[Respond to this article](#)

<http://rsta.royalsocietypublishing.org/letters/submit/roypta;368/1925/3813>

Subject collections

Articles on similar topics can be found in the following collections

[e-science](#) (31 articles)

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](#)

To subscribe to *Phil. Trans. R. Soc. A* go to:

<http://rsta.royalsocietypublishing.org/subscriptions>

Deploying general-purpose virtual research environments for humanities research

BY TOBIAS BLANKE¹, LEONARDO CANDELA², MARK HEDGES^{1,*},
MIKE PRIDDY¹ AND FABIO SIMEONI³

¹*King's College London, Centre for e-Research, 26–29 Drury Lane,
London WC2B 5RL, UK*

²*Istituto di Scienza e Tecnologie dell'Informazione ISTI-CNR, Via G.
Moruzzi 1, 56124 Pisa, Italy*

³*University of Strathclyde, Department of Computing Science,
26 Richmond Street, Glasgow G1 1XH, UK*

Several virtual research environments (VREs) have been developed to address specific tasks or application domains. Building on the experiences and use cases coming out of these projects, this paper addresses the creation of more general-purpose VREs for the humanities, which move beyond specific, focused tasks, and instead provide services and environments that support more general-purpose humanities research activities. Specifically, we are investigating use cases related to the organization and integration of the dispersed and heterogeneous information on which such research is based. These use cases are highly interactive, interpretative and researcher centric, addressing topics such as annotation environments and support for 'active-reading' processes and scholarly dialogues. We present the background to our work and the technical approach taken, and report the results obtained so far.

Keywords: humanities; virtual research environments; gCube; digital archives

1. Introduction

Many specialized virtual research environments (VREs; Fraser 2005) have been developed to address particular tasks in various humanities disciplines. By VRE, we understand a collaborative digital environment that facilitates the integration of information resources and tools for supporting research activities. 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/>) develops services for sharing and annotating manuscripts. The King's College London-based TEXTvre (<http://textvre.cerch.kcl.ac.uk/>) project is concerned with the institutional integration of VREs in the specialized domain of digital humanities, specifically the creation of Extensible Markup Language (XML)-based resources.

*Author for correspondence (mark.hedges@kcl.ac.uk)

One contribution of 16 to a Theme Issue 'e-Science: past, present and future I'.

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 organizing the heterogeneous and often unstructured digital resources 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 delivery of VREs to various humanities research communities in Europe. This paper describes experiments carried out to this end as part of the European Strategy for Research Infrastructures (ESFRI) project Digital Research Infrastructure for the Arts and Humanities (DARIAH; <http://www.dariah.eu>), which aims to conceptualize and build a virtual bridge between humanities and arts resources across Europe. Subsequently, we received funding under the Joint Information Systems Committee (JISC) VRE Rapid Innovation programme for the gMan project, which is consolidating these experiments.

The paper is organized as follows: in §2, we outline the previous research that led to the work described here, and in §3, we look at a more general motivation for developing general-purpose VREs for the humanities, in particular the importance of archives in humanities research. In §4, we introduce the grid-based technology framework used for our implementation. Section 5 describes the datasets used as the specific context for the use cases discussed in §6. Our experiments are outlined and discussed in §7, with some general considerations about the potential impact of the project on humanities research activities.

2. Previous work

The DARIAH experiments and gMan are based on use cases that were identified during the earlier research activities of the JISC ENGAGE project Linking and Querying Ancient Texts (LaQuAT; <http://laquat.cerch.kcl.ac.uk/>; Jackson *et al.* 2009). LaQuAT investigated how to integrate scattered, heterogeneous and autonomous data resources relating to ancient texts, mainly databases but also including XML documents (see §5). These resources were produced by various researchers in the Classics, and to varying degrees of quality as regards structuring of information. The assumption behind LaQuAT was that, although the development of standards such as Epigraphic Documents (EpiDoc; a set of recommendations for XML mark-up of inscriptions; Bodard 2008) is an important step forward in data interoperability, standardization is unlikely to solve all issues raised in linking up humanities data, for several reasons (Hedges *et al.* 2009). Firstly, there is a great deal of legacy data in diverse and often obsolete formats; secondly, training users in the application of a standard may incur a significant investment of time and money, which is not always available; finally, standards are generally developed within particular disciplines or domains, such as inscriptions, whereas research is often inter-disciplinary, making use of varied materials, and incorporating data conforming to different standards. There will inevitably be diversity of representation when information is gathered together from different domains and for different purposes, and consequently there will always be a need to integrate this diversity.

LaQuAT attempted to solve these issues by offering a flexible data integration framework based on workflows and the Open Grid Service Infrastructure-Data Access Integration (OGSA-DAI; <http://www.ogsadai.org.uk/>) grid middleware. OGSA-DAI provides a set of query, transformation and delivery interfaces that support virtualization of diverse data resources, primarily relational databases, but also XML-based resources and (by developing additional modules) other data. It thus allowed us to provide the researcher with integrated virtual interfaces to the various data resources that they need to access.

LaQuAT's results were useful from both a humanities and a computer science perspective. On the one hand, humanities researchers were able to open up new lines of enquiry by combining existing data resources, for example by discovering references to homonymous (and possibly identical) persons in different texts that could be dated to within a small number of years of one another. On the other hand, the many problems raised by integrating the human-created resources that are common to humanities research led to the development of the OGSA-DAI framework in new directions.

Nevertheless, LaQuAT also identified limitations to this approach to data integration in the case of humanities resources. In general terms, OGSA-DAI is optimized for working with *data-centric* resources rather than *text-centric* resources. The distinction here is between resources that contain significant quantities of unstructured text (text centric), and those that consist primarily of structured data such as numerical data, dates or very short text fields. Indeed, the limitations of the approach became particularly apparent when it came to working with XML files of inscriptions rather than with databases (Jackson *et al.* 2009).

In the humanities, researchers work more commonly with text-centric XML, essentially text documents, marked up as XML to capture document structure and some additional metadata (Nentwich 2003). Here, it is often more important to find sufficient relevant information in the texts, so that standard document retrieval techniques can be applied and adopted to deal with the specifics of handling additional structural constraints (Blanke *et al.* 2007).

As we saw in LaQuAT, databases are also used in the humanities to manage long text fields, whether because database technologies were easily available and understandable, or because researchers did not have the resources needed for an XML-based approach. Although relational databases have recently added text-search features, these are often insufficient for the complex tasks required in humanities research. Moreover, OGSA-DAI currently has no mechanism for promoting database text indexes to the level of the overall virtual database (although other indexes *are* promoted), which makes it impossible to search and retrieve lists of textual resources across various databases, ranked by how relevant these resources are to a user's specific research need.

Thus, OGSA-DAI places limitations on the sort of work that a humanities researcher is able to do. It works well if the structural context of the information is clear and the query aims at exact matches, for example, finding Roman Emperors who are mentioned in the inscriptions of Aphrodisias. Such database-style queries require no ranking of result lists. Most often, however, humanities researchers are looking for resources for further reading. They would like to be presented with a set of resources that are organized by relevance to their research need, e.g. if they

search for all Roman legal texts in one data resource that contain information on punishments for murder, which are also mentioned in papyrus documents from another data resource.

Related problems arise when dealing with humanities datasets in general (Blanke *et al.* 2008), problems stemming directly from the semantics of integrating information that is incomplete, uncertain and inconsistent, issues to which database technologies are particularly sensitive. For example, LaQuAT investigated *join queries*, to use Structured Query Language (SQL)-like terminology, that is queries across more than one database that filter the result set by joining columns in different databases, but concluded that the generation of meaningful links between humanities data resources in such an automated fashion was highly problematic, as the semantics of the relationships between different resources were unclear.

These conclusions from LaQuAT were further elaborated in the use cases that were developed from them as part of the DARIAH project (see §6), and are the main drivers for the work described in this paper. Complementing this is a significant body of methodological investigation centred around humanities researchers and their use of sources, particularly concerning the use of data and archives. Before describing our current work, we will survey briefly these investigations.

3. Data and archives in humanities research

The difference in scholarly practices between the sciences and mainstream humanities is highlighted in a study (Palmer *et al.* 2009) that 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 are 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 are 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, namely in archives, and it is in archives that the scholar carries out the work of assessing these source materials. Archival records are primary sources about the past and may take many forms, such as government correspondence, financial documents, photographs, sound recordings, etc. All this information is unstructured in its nature and is accessible via finding aids, which themselves are not structured information, but are, in most cases, documents containing

detailed information about the records in a specific archival collection. They are the primary source of information for researchers for assessing the relevance of a collection.

For instance in the UK, the National Archive preserves government records, while the National Monuments Record is the public archive of English Heritage, and there are similar institutions across Scotland, Wales and Northern Ireland. Furthermore, there is a plethora of local archives, many of which have contributed to the A2A programme (<http://www.nationalarchives.gov.uk/A2A/>), which makes it possible to search across collections online. However, even with online finding aids such as these, it is difficult to locate information in archives. Archival material can be found in many formats and places, and Archival Information Systems are often non-intuitive to use. In Duff *et al.* (2004), geographic location and lack of finding aids are seen as primary barriers to access. It is also often impossible to find things that are not already known because they are uncatalogued, and few finding aids actually help with the content of the records, but concentrate on their description (i.e. metadata).

In the UK and elsewhere, there are significant digitization programmes for archival material that 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 digitizing the material for which they are responsible, but nevertheless digitization is on the whole a somewhat piecemeal affair, and is carried out to different extents (e.g. image only or image plus optical character recognition) 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) archival content can be delivered to humanities researchers more effectively, independently of the location and implementation of that content, and with special means provided for customizing 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 specialized 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.

4. Technology framework

LaQuAT and the VREs described in §1 each pioneer particular opportunities for collaborative, data-driven research in the humanities. However, our investigations have demonstrated a need among humanities researchers for more general-purpose, scalable and cost-effective approaches to managing and manipulating data that go beyond the ad hoc integration of a small number of data sources.

Here, the humanities are no different from the sciences in requiring mechanisms, standards and policies for the controlled and dynamic sharing of hardware, software and data resources across organizational and disciplinary boundaries.

This is the Grid vision pursued by many e-infrastructure initiatives for science, some of which provide direct support for building VREs from infrastructural resources. 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. *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 (Candela *et al.* 2008). To this end, gCube interfaces with European grid middleware and research infrastructures (Enabling Grid for e-Science and the future European Grid Initiative) to exploit shared access to computational and storage resources, complementing this with an array of services that collate, import, describe, annotate, merge, transform, index, search and present information for various multi-disciplinary communities. These higher-level services are distributed functionally across three layers, as shown in figure 1.

The particular appeal of this approach is in its integration and transparency: services, information and machines are infrastructural resources that communities select, share, compose and consume in the scope of a VRE. VREs are 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.

We are investigating how humanities data sources 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. In the following sections, we describe the datasets and use cases that inform our development of new gCube services, which we use in the experiments described in §7.

5. Test datasets

As initial test datasets for our experimental scenarios, we are using the following three resources from LaQuAT:

- (1) The Heidelberger Gesamtverzeichnis (HGV) der griechischen Papyruskunden Ägyptens (<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.

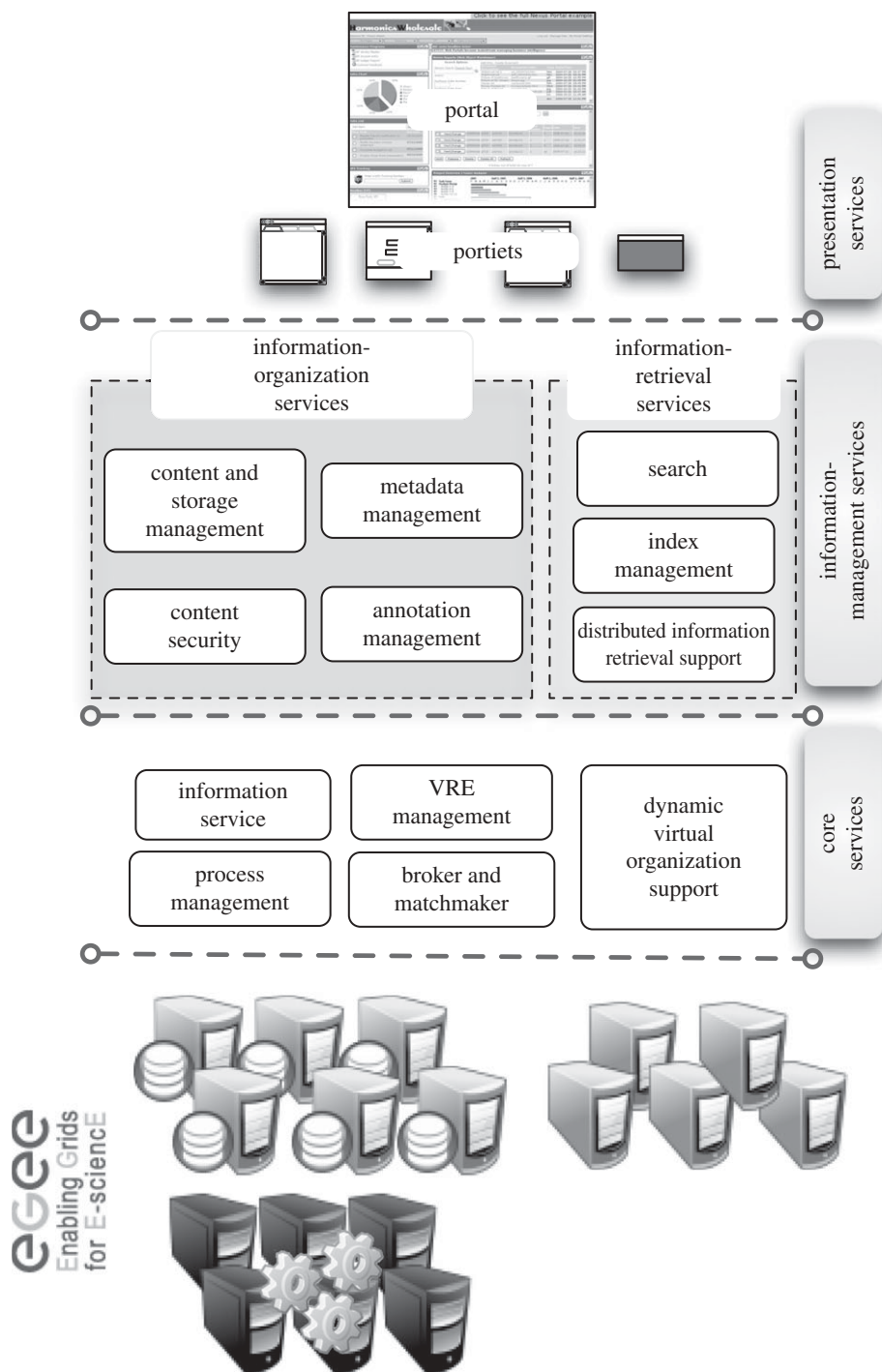


Figure 1. gCube architecture.

- (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 2000 ancient Greek inscriptions from the Roman city of Aphrodisias in Asia Minor, including transcribed texts and metadata marked up using EpiDoc Text Encoding Initiative (TEI), as well as images of the physical objects.

We are supplementing these with three online resources that are essentially collections of ‘things’—respectively, places, personal names and coins—each of which is identified by a stable Uniform Resource Locator (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, JavaScript Object Notation (JSON) or Resource Description Framework (RDF).
- (6) The *American Numismatic Society’s* collection of coins, whose entries can be referenced by URLs that return HyperText Markup Language (HTML), or by Domain Name (Space) Identifiers (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, and re-using them will facilitate a critical comparison of results with those from LaQuAT. The supplementary resources provide useful domains for annotations and inter-object links, as there are numerous potential connections with the first three datasets.

6. Humanities use cases

The outcomes of LaQuAT demonstrated the feasibility of and need for environments in which ancient historians can access and manipulate dispersed and heterogeneous digital resources in an integrated way, and these requirements were further elaborated in use cases produced by DARIAH. These use cases were derived by engaging with researchers at King’s College London through semi-structured interviews, in which the questions were organized around viewpoints and concerns that examined how work is coordinated, how it is planned and formalized, how computer and paper processes are used together and how physical space and time impact on the work.

These use cases gave rise to a view of research activities in this discipline as being complex and highly interactive workflows, with the researcher at the centre. The researcher looks for resources, which may be text or data centric, relevant to her interests, selects, interprets and analyses them, using tools but also her own judgement based on other evidence, both internal and external to the resource. The results of one search may, taken together with other information available to a researcher, influence the questions that are asked of others. Importantly, this may be a collaborative activity, requiring the ability to record interpretation and opinion as annotations, and thus start a dialogue within the relevant community. A key issue, and one that drives the approach we are taking, is that the researcher requires an approach that is more text centric than was possible in LaQuAT. All the LaQuAT databases contain documents such as Roman law texts. Most likely, the researcher will be interested in these.

The use cases can thus be very varied and unpredictable. Our approach to developing support for this research community was to break the use cases identified in interviews down into a number of common, atomic actions. Specific instances of these actions can be combined to model a variety of ‘real’ research scenarios.

- The researcher assembles heterogeneous resources (or parts of resources) into a virtual collection, either manually (like a shopping basket) or by specifying membership criteria. By *virtual collection*, we understand a set of resources that the researcher can manipulate as if it were an object in itself, e.g. refer to it by name/identifier, search across it and share it with colleagues. The ability to build such virtual collections, while conceptually straightforward, is very important in allowing researchers to deal with the large quantity of archival documents and information with which they have to deal.
- The researcher performs a text-centric search across a virtual collection. By text-centric search, we understand a search across documents that copes with uncertainty and non-exact matching, and produces a ranked result set, analogous to how a search engine deals with the Web. In this way, the researcher is helped to find relevant resources, filter and select from them, and use them as the basis for further searches.
- The researcher performs a date-based query across a virtual collection. This is particularly challenging in our context, for several reasons. Dates are represented in ancient documents in various ways, which are not easy to compare or map into modern terminology. Dates may be proposed by researchers on the basis of, say, writing style or archaeological evidence, and may be subject to different degrees of uncertainty. The precision of temporal data may vary, from a specific day to a century or more. Date-based queries may involve searches across both structured and unstructured resources (e.g. databases and texts).
- The researcher performs a geo-spatial search across a virtual collection. This is again a challenge, not only is location given in a variety of different ways, with wide variations in precision and accuracy, such searches can involve highly diverse types of resource (e.g. databases and maps).

- The researcher annotates a research object (e.g. an XML file, an image, a row in a database), or part of an object (e.g. a word within an XML file, a position or area within an image, a cell within a database record). For example, the researcher may add additional information, such as the identity of a person mentioned in a text, or an explanation of the usage of an unfamiliar word; in contrast, the researcher may consider existing information to be erroneous—for example the supposed date of a papyrus, or the transcription of damaged text—and so add an annotation indicating the reasons for this conclusion.
- The researcher adds a link between two research objects, or parts of objects. The LaQuAT project concluded that it is difficult to generate meaningful links between resources in a purely automated fashion because of the uncertainty of much of the data, and that researchers form such connections by their own judgement. For example, she may decide that a papyrus in one archive is related to a papyrus in another—perhaps they are in the same distinctive handwriting, or refer to business transactions by the same merchant. Such connections are by no means certain, and the researcher wants not just to be able to indicate that ‘pap1 hasSameScribeAs pap2’, but to add justification for this inference.
- The researcher can search across annotations and links as well as the ‘original’ resources.
- While it is the researcher who makes the decision when creating an annotation or link, it would be useful if they could be generated in a partly automated way, i.e. by the system providing suggestions that are verified (or modified) by the researcher. For example, a query to connect up two data resources with reference to a particular set of criteria, for example a date range, might first match entries in the original data resources and then search the annotations, and propose a list of possible connections.
- The researcher shares her work, including the relevant research material, annotations and links, with selected colleagues, who then in turn add their own annotations and links that may confirm, extend or contest the researcher’s conclusions. In this way, a scholarly dialogue is created and recorded. This could also facilitate new forms of publishing in the humanities, in which readers have access to the reasoning process that lies behind conclusions, enabling them to validate it, and perhaps criticize it. As observed in §2, humanities research often depends on provenance of information and peer assessment.

7. Experiments

(a) Overview

The tools provided by gCube extend the opportunities for the management and manipulation of humanities datasets beyond those provided by the data-centric OGSA-DAI used in LaQuAT, particularly in terms of annotation, reporting, sharing and text-centric queries. Prior to investigating these opportunities, however, we endeavoured to reproduce, within the context of gCube, queries and results analogous to those investigated during LaQuAT. The queries

investigated were all driven by the use cases obtained from researchers, so they corresponded to questions that researchers might want to ask of their datasets in order to develop answers to genuine research questions in the field of ancient history. Indeed, they were based on a series of interviews with a group of such researchers.

(b) *Importing data into gCube*

First note that, although database resources can be imported into the gCube environment, gCube uses its own generic representation of data. The database model is based on tables and rows, whereas that in gCube is based on documents and collections of documents, where a document may be a structured, compound object comprising multiple, nested components, with multiple representations, metadata and annotations. This data model is implemented in terms of atomic information objects and typed relationships between those objects, i.e. a compound document is implemented as an annotated graph with information objects as nodes. Thus, a database is not simply imported into gCube in a single, standardized manner; instead, different mapping strategies can be specified (and implemented as import scripts) that define how the information represented in the databases is modelled using information objects within gCube. So, for example, individual tables, individual rows or even individual cells could be mapped onto separate information objects. This is quite different from OGSA-DAI, which retains the relational model followed by the original databases.

Secondly, note that importing a resource need not mean that the actual physical content of the resource is stored within and managed by the gCube infrastructure, but rather that the logical structure of the original resource is described within gCube in accordance with its information object data model. The actual physical content may continue to be held elsewhere, so long as it remains accessible.

Although the two database resources that we used both comprise several tables, in each case, the database describes directly a collection of documents (in the archival sense of document): papyri in the case of HGV, Roman law texts in the case of Volterra. In the former case, the full texts were not included in the database, although many of the records contain identifiers to other corpora from which the full text could be extracted. As these documents are the intellectual entities with which the researcher is dealing, the natural mapping is for each document described in the databases to correspond to one Information Object in gCube. The correspondence between documents and rows is not quite one-to-one, however, as in a few cases, a single papyrus may correspond to more than one HGV entry. Such multiple entries are linked by the HGV identifier, and for simplicity they were preserved by the mapping into gCube.

Basic metadata associated with the Information Objects was extracted from the corresponding database records, and represented as Dublin Core, e.g. dc:coverage, including geographical and temporal attributes, and dc:subject, including keywords. InsAph is XML-based, comprising XML files corresponding to individual inscriptions, with associated metadata and images. Again, the natural mapping for the scholar was for a single inscription to correspond to a single Information Object. We have not imported the remaining resources (4–6 in §5), but rather are using them as a domain for annotations. This is a realistic

scenario, as given the variety of ancillary web-based reference resources that a user might need, such as dictionaries, prosopographies or concordances, it may not be realistic to import all of them.

(c) *Consolidation of LaQuAT experiments within gCube*

Our initial experiments aimed to reproduce within gCube queries analogous to those investigated during the LaQuAT project. These queries addressed the following broad lines of enquiry:

- Investigating the chronology of events at a particular location, for example by interrogating the papyrus records from a particular origin (from HGV), together with the legal texts for that same location (from Volterra) with particular reference to the date or date range.
- Investigating the patterns of activities of individuals or groups of individuals in a particular social milieu, and the patterns of relationships between them, for example by retrieving records (papyri, law texts or inscriptions) that contain references to certain individuals, or, more precisely, that contain certain names. As observed above, the identity of the referenced individuals is a matter of judgement for researchers.
- Investigating the occurrence of particular types of event and activity during certain periods and in certain places, for example, events relating to civic or personal life, by interrogating the documents for words associated with such events. An archaeologist, for example, might want to correlate such documentary traces with evidence from the material record, such as excavation reports.

In SQL-like terminology, these might be described as *union queries*, by which we mean, in this context, queries that search across, and aggregate results from, multiple databases. To save space, we will only look in detail at specific queries relating to (1) above: figure 2 represents three such queries using SQL-like notation for clarity, together with a summary of the result sets in each case. It is clear that, in numerical terms, the overlaps between these datasets are not very large. The point of the exercise, however, was to demonstrate the principle; the resources used were just examples taken from a much wider digital environment, and the practical efficacy of these scenarios will increase significantly once a certain ‘critical mass’ is reached.

As expected, using the data mapping described above, we produced in gCube results analogous to those obtained in LaQuAT, but with the difference that the results were now relevance ranked. The next step was to address the requirements outlined in §6.

(d) *Implementing a Classics research scenario within gCube*

The possibilities offered by gCube are very promising—they are described from a user perspective in the user guide—and indeed the close correspondence of gCube’s functionality to our use cases is one reason that we are investigating it. This work is still ongoing, and we will report later on a full and systematic set of experiments aimed at supporting these requirements. Here, we give an overview

```

Select all records from Volterra_DB and HGV_DB
where Volterra_DB.DatLocation(recorded), HGV_DB.ort LIKE '%
  Antiochia%'
and Volterra_DB.datum (preferred) between "210/01/01" AND "
  260/01/01", HGV_DB.J between "210" AND "260";
Result: This returns 1 Volterra record and 1 HGV record.
Select all records from Volterra_DB and HGV_DB
where Volterra_DB.DatLocation(recorded), HGV_DB.ort LIKE
  '%Apollonopolis%'
and Volterra_DB.datum (preferred) between "200/*/*" AND "400/*/*"
  , HGV_DB.J between "200" AND "400";
Result: This returns 1 Volterra record and 5 HGV records with
  year or J (215,217,227,243,236).
Select all records from Volterra_DB and HGV_DB
where Volterra_DB.DatLocation(recorded), HGV_DB.ort ="Alexandria"
and HGV_DB.J between "200" AND "400";
Result: This returns 1 Volterra record and 86 HGV records.

```

Figure 2. Example queries.

of our initial experiments, in which we address a research scenario involving the following stages:

- document-centric and text-centric search; creation of virtual collections,
- creation of annotations and links, and
- generation of research reports.

gCube incorporates a variety of search approaches. It also supports the creation of *Collection* objects, which may be defined either explicitly by lists, or implicitly by membership criteria, and may be structured by nesting. These objects fulfil the role of our virtual collections. The following operations were performed:

- create collections for the result sets from §7c and
- create a collection of papyri, inscriptions and law texts relating to civic activities. This was not just a simple search, but involved multiple text searches on various terms, filtering out of irrelevant matches and combining into a structured collection.

The gCube system supports annotation and linking via the creation of *Annotation* and *Association* objects. The former of these allow a textual note to be attached to an object in gCube, marked with the timestamp and the user who created it. The latter allow labelled links to be created between objects. In addition, for text-based and image objects, Annotations can be added to parts of an object. The following operations were performed:

- select the Volterra record containing the text ‘Antiochia’, retrieved above, and add an Annotation ‘This is Antioch by the Euphrates, Pleiades reference <http://pleiades.stoa.org/places/658562/antiochia-ad-euphratem/>’; finally, create an Association with the Pleiades object,

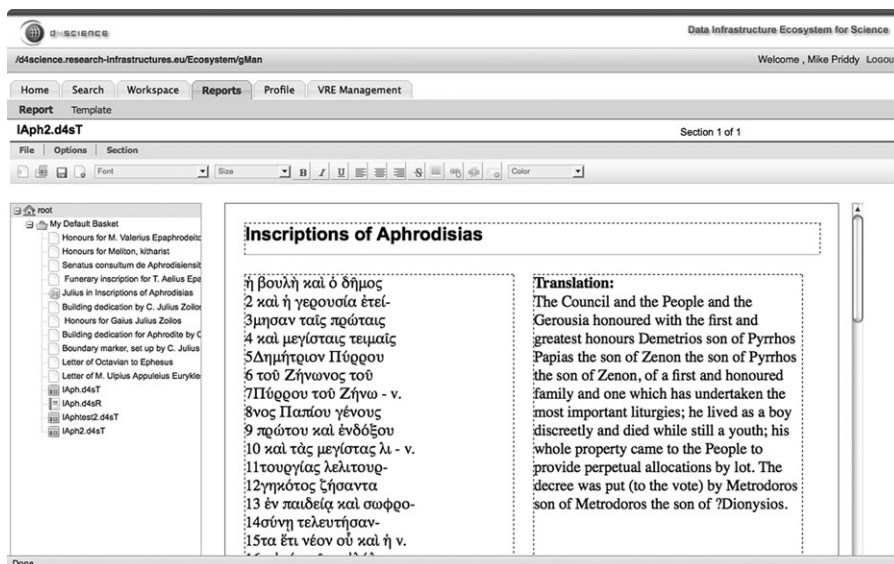


Figure 3. Report on inscriptions of Aphrodisias.

- select a papyrus from the above collection, select a transcribed personal name in the text, and add an Annotation: ‘This name should be XXXX, not YYYY. XXXX is attested in other papyri from the same location and time’,
- select the corresponding image file, select the region containing the name and add the same annotation, and
- select two papyri from Alexandria and add an Association tagged ‘isFromSameLocation’.

Finally, gCube allows reports to be created based on results retrieved from its environment. This typically proceeds in two stages: firstly, a report template is defined, representing the report’s structure and containing static information; secondly, the report itself is created by exporting the content into the template. The screenshot in figure 3 shows a simple example of this procedure performed during the scenario described above. These reports allow researchers to describe and summarize their key data, inferences and (tentative) conclusions so as to facilitate communication and discussion of their work—as observed in §6, dialogue with colleagues (who may be dispersed internationally) is a key component of the scholarly process in much humanities research. The gCube reporting functionality also opens up possibilities for enhanced modes of scholarly communication that support publication of data as well as of articles (Woutersen-Windhower *et al.* 2009).

8. Conclusions and future work

In this paper, we have outlined how to implement a VRE for the humanities that, in contrast to much of the current VRE landscape, aims to support general-purpose research activities rather than particular, domain-focused tasks.

Specifically, we identified use cases in which the researcher organizes, annotates and integrates the largely unstructured digital resources used in humanities research, specifically in the field of ancient history and in the context of work in archives. Subsequently, as part of DARIAH, we performed experiments towards supporting these use cases.

These experiments demonstrated the feasibility of our approach, and with the success of these, we were encouraged to extend and consolidate this work. The continuing activities of gMan are developing more detailed use cases and providing a systematic evaluation of the gCube environment with reference to these use cases. To begin with, these experiments focused on individual actions and short scenarios, as outlined in §6. Subsequently, we carried out a number of more realistic and increasingly complex scenarios, which were based on our interviews with researchers and were representative of research activities in this field. One particular area of investigation for future work is the use of the gCube reporting functionality for implementing enhanced modes of scholarly communication that incorporate data publication. This would make it possible to use the environment to cover a very large part of the research life cycle subsequent to the creation of the primary datasets or archives, from ingestion of these datasets through to publication of research outputs.

The broader aim of our work is to develop and evaluate a means of providing general-purpose VREs for research communities in a variety of humanities domains, particularly those involved in archival work. The gMan project will provide a solid basis for this by allowing us to roll out an environment that exploits National Grid infrastructures and can be used and evaluated by humanities researchers in the UK and, via the DARIAH community, across Europe.

References

- Blanke, T., Hedges, M. & Dunn, S. 2007 E-science in the arts and humanities—from early experimentation to systematic investigation. In *E-Science '07: Proc. of 3rd IEEE Int. Conf. on e-Science and Grid Computing*. Washington, DC: IEEE Computer Society.
- Blanke, T., Aschenbrenner, A., Küster, M. & Ludwig, C. 2008 No claims for universal solutions. In *E-Science '08: Proc. of IEEE e-Humanities Workshop*. Washington, DC: IEEE Computer Society.
- Bodard, G. 2008 The inscriptions of aphrodisias as electronic publication: a user's perspective and a proposed paradigm. *Digital Medievalist* **4**, 22–27.
- Brockman, W., Newmann, L., Palmer, C. & Tidline, T. 2001 *Scholarly work in the humanities and the evolving information environment*. Washington, DC: Digital Library Federation, Council on Library and Information Resources.
- Candela, L., Castelli, D. & Pagano, P. 2008 gCube: a service-oriented application framework on the grid. *ERCIM News* **72**, 48–49.
- Candela, L., Castelli, D. & Pagano, P. 2009 On-demand virtual research environments and the changing roles of librarians. *Libr. Hi Tech* **27**, 239–251. (doi:10.1108/07378830910968191)
- Duff, W., Craig, B. & Cherry, J. 2004 Historians use of archival sources: promises and pitfalls of the digital age. *The Public Historian* **26**, 7–22. (doi:10.1525/tph.2004.26.2.7)
- Fraser, M. 2005 Virtual research environments: overview and activity. *Ariadne* **44**, 31–40.
- Hedges, M., Blanke, T. & Hasan, A. 2009 Rule-based curation and preservation of data. *FGCS, Future Gener. Comput. Syst.* **25**, 446–452. (doi:10.1016/j.future.2008.10.003)
- Jackson, M., Antonioletti, M., Blanke, T., Bodard, G., Hedges, M., Hume, A. & Rajbhandari, S. 2009 Building bridges between islands of data—an investigation into distributed

- datamanagement in the humanities. In *Proc. 5th IEEE Int. Conf. on e-Science*. Washington, DC: IEEE Computer Society.
- Nentwich, M. 2003 *Cyberscience: research in the age of the internet*. Vienna, Austria: Austrian Academy of Sciences Press.
- 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. See <http://www.oclc.org/programs/publications/reports/2009-02.pdf>.
- Woutersen-Windhouver, S., Brandsma, R., Hogenaar, A., Hoogerwerf, M., Doorenbosch, P., Dürr, E., Ludwig, J., Schmidt, B. & Sierman, B. 2009 *Enhanced publications: linking publications and research data in digital repositories*. Amsterdam, The Netherlands: Amsterdam University Press.