LaQuAT: Integrating and Querying Diverse Digital Resources in Classical Epigraphy

Tobias Blanke,[1] Gabriel Bodard,[2] Stuart Dunn,[1] Mark Hedges,[1] Michael Jackson,[3] Shrija Rajbhandari[1]

[1] Centre for e-Research, King's College London. United Kingdom.
[2] Centre for Computing in the Humanities, King's College London. United Kingdom.
[3] EPCC, University of Edinburgh. United Kingdom.

Abstract
Databases and corpora of marked-up texts form a significant proportion of the outputs of digital archaeology. Although the development of standards for the representation of such information creates new possibilities for interoperability, significant problems remain. Legacy data exists in diverse and often obsolete formats, and even when standards are applied, the sheer variety of cultural data and research means that there remains a great deal of 'fuzziness'. Users must be trained in the correct application of standards, entailing significant outlay of time and money. Moreover, there is inevitably diversity of representation when information is gathered together from different projects and for different purposes, and thus there will always be a need to integrate this diversity while preserving the integrity of the data itself.

The LaQuAT (Linking and Querying Ancient Texts) project investigated technologies for providing integrated SQL-based views of diverse data resources related to classical archaeology, specifically containing epigraphic and papyrological material. These resources are quite heterogeneous in terms of standards and structure, comprising two relational databases with different schemas, and an XML-based corpus; they are hosted by different institutions in different countries, and are the outputs of divergent research communities. Nevertheless, the datasets overlapped geographically, chronologically and prosopographically. Such issues are typical of archaeological databases; to address them the project explored the applicability of 'grid computing', and in particular the OGSA-DAI software, for providing integrated views of such diversity.

Keywords: *epigraphy, papyrology, data integration, OGSA-DAI, grid*

[A]rchaeologists should look beyond the short term when planning how to use a computer. The world of archaeology is likely to be considerably different in twenty years from now (2009), so archaeologists need to plan with future change in mind.[1]

1 INTRODUCTION

The massive growth in the quantity of archaeological data, along with the development of the Internet and ubiquitous online research networks, have undoubtedly proved Moffett correct in his prediction. Increasingly sophisticated data capture techniques, and wider uptake of them, are resulting in a large volume of complex data being held in databases, which need to be accessible if they are to be useful. This in turn places requirements on database technology that are characteristic of archaeology.[2] One key requirement, which has received much recognition in recent literature, is the need to integrate data from heterogeneous sources.[3] It would be difficult to understand a single floor level, or pottery deposit, or set of post holes without reference to other comparable data, be it from the immediate vicinity, from the same site, from an intra-site or even regional context. Archaeological databases need to be treated in a similar fashion: in order to be useful for making valid archaeological or historical interpretations, they must be integrated with each other, wherever and however they were created. This paper describes the Linking and Querying Ancient Texts (LaQuAT) project, funded under the JISC ENGAGE Initiative (http://engage.ac.uk/engage). LaQuAT was a collaboration between the Centre for e-Research (CeRch) and the Centre for Computing in the Humanities (CCH), both at King's College London, and the Edinburgh Parallel Computing Centre (EPCC) at the University of Edinburgh. The project's aim was to explore 'grid computing' technologies as a means of achieving such integration, using as a demonstrator three separate datasets containing Roman legal and epigraphic texts.

Any attempt to describe the archaeological research cycle by way of an introduction to a paper about databases inevitably risks over-generalization. However, it is important to gain a sense of where in the

[1] Jonathan Moffett, "Computers in Archaeology: Approaches and Applications Past and Present," in *Computers for Archaeologists*, ed. Seamus Ross, Jonathan Moffett and James Henderson (Oxford: Oxford University Committee for Archaeology Monograph No. 18, 1991) 18.

[2] Gary Lock, *Using Computers in Archaeology: Towards Virtual Pasts* (London: Routledge, 2003) 90–98.

[3] E.g. Dean R. Snow, et al., "Cybertools and Archaeology," *Science* 311 (2006): 958–959.

archaeological process such a project lies. The research cycle, leading from the discovery and excavation of archaeological material in the field to its publication, contains a number of discrete elements, which can be generalized as follows:

- *Discovery*. An artifact is recovered from the field, either through survey or excavation.
- *Identification and attribution*. One or more attributes, including (but not necessarily exhaustively) object class, type, color, dimensions and a provisional date are assigned to the artifact, based on its physical characteristics.
- *Reference and cross-referencing*. The artifact is compared, whether using some formal measure such as dimensions or color; or interpretively, where the researcher attributes it to a particular group or typology, with artifacts of similar type.
- *Interpretation*. The artifact's place in the wider regional and spatial context is determined.
- *Publication and archiving*. Representations and/or textual descriptions of the artifact are published either electronically or on paper. Electronic publications may be deposited in a digital repository. The artifact itself may enter a museum collection or other archive, where it will be given a non-random place within a context of other artifacts sharing its attributes.

Because archaeological fieldwork is by definition regional or site-specific, many excavation activities generally focus their efforts at any one time on relatively small-scale data gathering activities. This produces bodies of data that might be archaeologically comparable, but are not recorded or represented consistently. Furthermore, much archaeology in the UK and elsewhere is conducted in order to fulfilll the legal obligations of land developers, which limits the time, human and financial resources available for excavation; moreover, most organizations that carry out such 'rescue' excavations—university archaeological service units, local authorities, private consultancies—have their own recording systems and procedures. It follows that there are still fewer resources available for the effective preservation and curation of complex information arising from such excavations.

The problem of unique, independent data silos can be found in many research disciplines. In many areas of the so-called digital humanities, we will find data resources that are created as the result of a particular project's focus on digitizing a collection of medieval documents, modern newspapers, or other resources for historic research. None of these collections will have been created with the perspective of integrating them and bringing the information in them together, so that in combination this information could become more than the sum of the individual information items. Without doubt, however, linking, for example, the finds of excavations of ancient Roman towns will help archaeological research. The aim of the LaQuAT project was to showcase how this could be done, but also to identify the challenges involved in attempting to do so. These challenges are not only ones of technology but also of understanding this new way of looking at resources in the humanities in general and in archaeology in particular.

2 GRID COMPUTING AND ARCHAEOLOGY

This paper focuses on how grid computing, which has proved successful for integrating data resources in the natural science disciplines, can aid the integration of comparable resources in archaeology. The web allows users all over the world to connect documents using protocols such as Hypertext Markup Language (HTML). The grid is a parallel architecture that enables not only documents, but also resources to be connected. Instead of a human searching for information, a *resource broker* will find appropriate resources for the computing task at hand. From the point of view of the LaQuAT project, the task would consist of connecting various independent data resources in such a way that a global virtual data space for archaeology results. Generally, the resource broker is the centerpiece of a grid, allowing it to connect resources around the world to become a distributed virtual computer, which is universally available to a research community.

LaQuAT uses the Open Grid Services Architecture Data Access and Integration (OGSA-DAI) project (www.ogsadai.org.uk/). OGSA-DAI is a set of middleware components that allows heterogeneous data resources to be connected using web services. It has been used extensively in the natural and physical sciences[1] (see www.ogsadai.org.uk/about/projects.php), but as far as we know, LaQuAT is the first application of OGSA-DAI in archaeology, or indeed in the humanities.

OGSA-DAI provides a framework for the access, management, and integration of distributed, heterogeneous data resources. It does so by adding standardized workflows of related processing to the data, and executing workflows that can be viewed as scripts specifying what data is to be accessed, and what is to be done to it. Workflows consist of activities, which are well-defined functional units that perform some data-related operation. This can include querying a database, transforming data to XML, or delivering data via FTP. A client submits a workflow to an OGSA-DAI server via an OGSA-DAI web service. The server parses, compiles, and executes the workflow. Depending upon the client's request, the data may then be returned from the service to the client. Using such workflows, we can therefore provide 'on-the-fly' common virtual interfaces to data. The OGSA-DAI project supports the exposure of data resources, such as relational or XML databases, on grids. Various interfaces are provided and many database management

[1] A. Grant et al., "OGSA-DAI: Middleware for Data Integration: Selected Applications." *IEEE Fourth International Conference on eScience, 7–12 Dec. 2008*:.343.

systems are supported, with a particular view to querying, transforming, and delivering data in different ways via a simple toolkit for developing client applications. OGSA-DAI is designed to be extensible, so users can provide their own additional functionality.

In order to facilitate information access across independent data resources, LaQuAT uses OGSA-DAI's service-based Distributed Query Processor (DQP), which is able to execute queries in parallel over OGSA-DAI services and other (Web) services. It combines data access with analysis. Simply put, the DQP package makes tables residing in multiple distributed databases appear to a user as if they are tables within a single database.

An important feature of OGSA-DAI is that it does not affect the original databases in any way. There is no need to define a new database with a schema that combines the original ones; rather, OGSA-DAI allows a virtual integration that respects the autonomy of the originals.

For LaQuAT, the latest version of OGSA-DAI, OGSA-DAI 3.1 (Axis), released in December 2008, was used. This version included some bug fixes that were identified in the early phases of the LaQuAT project, although additional bug fixes were required as the project progressed.[1]

3 THE PROJECT

LaQuAT uses this architecture to demonstrate how databases of documents from the Greco-Roman world can be linked. The texts include papyri from the Egyptian desert, inscriptions (on stone), and texts that survived by repeated copying. Inscriptions are a critical source of information for studying the ancient world, and for many years now, researchers in classics and related disciplines such as papyrology have been investigating, publishing, and commenting on these texts. This process has created a substantial body of digital material of one form or another—a lot of these objects are databases, others are texts marked up in various ways, more recent ones in XML, older ones in SGML. Furthermore, and in common with many other types of archaeological data, the formats of inscription databases are very diverse. It is unlikely that any two of the databases in question follow a common database schema, and the markup can vary wildly, particularly in older cases when less effort was made in standardization. Secondly, they are not generally available for use—in many cases they are locked away on departmental machines, while in other cases they are "published" on a web site, but not in a way that make the data particularly usable. A key aim must be to make data available in such a way that it can be processed (e.g. for data mining), rather than just browsed. Thirdly, even when they are available, they are published in isolation. Many of these resources may be regarded as fragments of a larger picture, and would have vastly more value if researchers could have access to this larger picture rather than just the parts.

A further factor is that resources may be owned by different communities and subject to different rights. The scholars who created them may be unwilling to accept anything that affects the integrity of the original resources, and may be reticent about publishing 'unfinished products'. Consequently, any integration initiative must respect the autonomy and integrity of any rights-holders if it is to gain acceptance.

The LaQuAT project addressed these issues by means of a demonstrator that incorporated databases with different schemas, as well as XML-based data, and that provided an integrated view of the data that was useful to researchers in the field. The project endeavoured to minimize any changes required to original datasets, and noted and investigated any issues that arose.

4 DATA SOURCES

As indicated above, there are many datasets to which such an approach could be applied. For the purposes of LaQuAT, we selected three datasets, which had a certain degree of spatial and temporal overlap, and which were easily accessed. These were:

Projet Volterra. This is a database of late Roman legal texts, housed at University College London (www.ucl.ac.uk/history2/volterra/index.htm). The project has worked to produce a database that contains the basic texts of imperial legal pronouncements (where the verbatim text survives) from any source, be it epigraphic, papyrological, juristic, or literary. A Microsoft Access database was created to contain the edited texts of all imperial pronouncements in Latin for the entire period A.D. 193–455, divided into six chronological tables (consisting a total of 5479 law records), with further supplementary tables. Microsoft Access was employed mainly because the epigraphers who created it were most familiar with this package. The structure of the database reflects this. The database contains texts and additional metadata about the texts, such as date, origin, etc. It contains ten tables, most of which contain individual laws. Although these were therefore conceptually similar, they were split up just to avoid having one big table. To add to the confusion, the columns vary from one table to the next.

The second dataset is the **Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV)**. HGV is a database of metadata for c. 65,000 papyri, including dates, people, and places (findspots, provenances), mostly from Roman Egypt and its environs. The database was made available over

[1]Mario Antonioletti et al. *Linking and Querying Ancient Texts (LaQuAT): Final Report*. Project report (forthcoming).

the internet by the Institute for Papyrology, Ruprecht-Karls-Universität Heidelberg. The project aims to provide a complete register of all documentary papyri (in Greek) from Egypt, and is therefore a highly valuable research tool. The project has built a database that stores the papyrological metadata with links to the electronic corpus of the published documents at The Duke Databank of Documentary Papyri (DDbDP). The database was created using FileMaker Pro, a proprietary system popular among humanities researchers because of its user-friendly data input function. However, it is not so easy to get it to interoperate with other database systems (nor with OGSA-DAI, as we discovered). The user-interface and the data itself are in German (available at: www.rzuser.uni-heidelberg.de/~gv0/Texte/HGV-Texte.html), although some help is available in English and Dutch. The FileMaker Pro database consists of a main table that contains all papyri, other supplementary tables for indexing and, "erwähnte Daten," listing the more than 9,000 papyri which contain explicit references to dates in the Greek text. The database is developed in accordance with the documentary publications of papyri, with a unique identifier for each papyrus.

The final dataset is the **Inscriptions of Aphrodisias**, a corpus of mainly Greek inscriptions from the city of Aphrodisias in Asia Minor (Turkey). The project publishes in electronic form the corpus of inscriptions relating to the period of the Roman Empire excavated from Aphrodisias. The Inscriptions archive (http:// in saph.kcl.ac.uk/iaph2007/) contains about 1,500 inscriptions, mainly on stone, found in the city or in its civic territory, up to the end of 1994. Transcriptions of the epigraphic text and archaeological data and context about the object are in EpiDoc (http://epidoc.sourceforge.net/) format.

LaQuAT implemented two case studies that would show how information in one of these data resources could be used to enhance and contextualize the information in the others. By conceptually linking at least two data resources, we aimed to show that new knowledge could be gained. Case Study 1 sought to integrate HGV and Projet Volterra. These two databases' content overlaps both chronologically and geographically; OGSA-DAI was employed with the relational views extension pack to produce a consistent schema between the two databases. The second case study sought to integrate the Projet Volterra database with the Inscriptions of Aphrodisias XML dataset. These datasets overlap both in the time period covered and in the people to which they refer.

Following the model of "data-driven development," the project formulated a number of hypothetical sample queries that archaeologists or epigraphers might wish to make. Data-driven development begins with users' data and information needs, and analyzes how users might use available resources to access services and fulfill those needs. These data and information needs can be formulated in various ways, depending on the users' level of expertise. Often they are simply natural language queries, but in this case we could expect the users, who are also the creators of the databases, to know how to formulate more advanced queries using SQL or XPath.

Two simple examples of such queries follow: (1) An historian of ancient social history might want to research the patterns of relationships and activities of individuals in a society of certain period. This research could be undertaken by analyzing the inscriptional data from the Aphrodisias database and the legal records of that period from Projet Volterra. In this case, he might want to locate all references in the legal records to a person named in the Aphrodisias inscriptions, during that time period. (2) A researcher may want to investigate the relationship between the application of laws in a particular place, and the official legal statements and regulations for that location during a particular period, by inspecting both the papyrus records and the official constitutional and legal pronouncement records of the time.

In SQL terms, these are *union queries*, queries to multiple databases that search across and deliver results from more than one field and across more than one database. The final aim of database integration however might be to produce *join queries*, combining, for example, fields from two tables by using values common to each. However, the existence of such common values cannot be assumed, and often they cannot be identified. Joining tables using non-unique values can lead to inconsistency.

5 ARCHITECTURE

Figure 1 shows how the LaQuAT architecture implements the integration of different database resources. OGSA-DQP is the main abstraction mechanism which will hide the details of a database implementation from the user. Our approach to virtual data integration is therefore to specify local data sources as views over the global schema.

OGSA-DAI uses SQL views to hide the details of a data resource. In OGSA-DAI, everything from a standard database, to an XML file, to an indexed text resource will look to the user as if he were interacting with a single large SQL data resource. To this end, OGSA-DAI generalizes the concept of an SQL view and virtualizes it. For LaQuAT, the following combination of traditional database technologies and OGSA-DAI technology will realize virtualization of data resources.

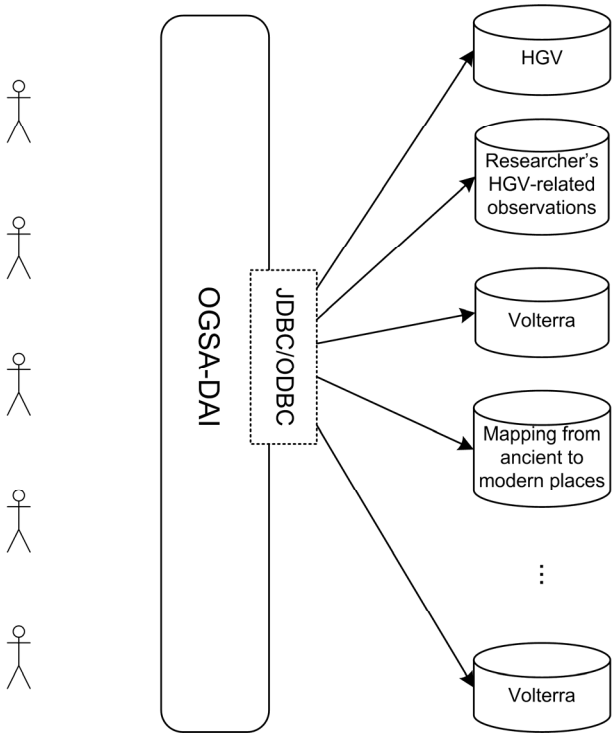

Figure 1*: Virtual Data Integration.*

In the case of the Volterra and HGV data resources, we also needed to bridge the language divide between the German HGV data resource and the Project Volterra resource, which is in English. This was achieved in the application by using a Join Table to map between German keywords and English keywords. This keyword mapping table was necessary only because of the language difference. In this table, we have two rows, one for the German word and one for the corresponding English word. We could have gone further in the integration of multilingual data resources by using DQP to hide the fact that HGV is a German database, so that the second data resource, Volterra, is not "aware" that it is interacting with a German resource. We decided to take the former approach, as we believe that other data resources might also benefit from having access to the translation table.

SQL views can handle the following requirements:

- *Expose TEXT date column types as DATE date column types. In Volterra, e.g., all date fields are defined as text fields in MS Access.*
- *UNION N tables so they are treated as a single table. This is standard view functionality, although some of the data resources have very specific ways of realizing them.*
- *Expose German column and table names as English.*

OGSA-DAI DQP can additionally handle the following requirements:

- *Expose multi-lingual column contents as English. This is done using the already mentioned Join table.*
- *Perform text searches over the contents of individual fields.*
- *Perform a join across databases.*

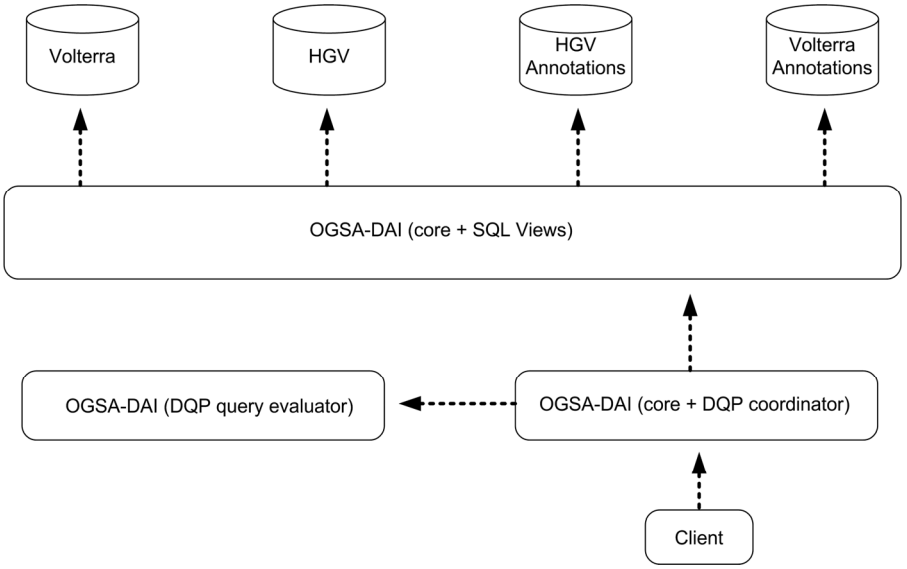

Figure 2: *LaQuAT Demonstrator Architecture.*

Our current design, based on the experiences and issues outlined in the initial LaQuAT experiments, is outlined in fig. 2. Instead of installing OGSA-DAI directly on the database server where the original data resource is located, we just use a JDBC connection behind OGSA-DAI and query the database from another dedicated server. The final goal would be a network of such servers, possible for multiple disciplines in the humanities, maintained by a trusted arts and humanities data service, perhaps along the lines of the former UK AHDS (http://www.ahds.ac.uk). In subsequent work, we plan to build a virtual data center that can integrate several archaeological data resources. The remote data source exposes its access to this data center, which could at the same time "monitor" the status of that remote data source. Each of the participating sites would agree in a contract with this data center to maintain an account to access the data (or alternatively a simple FTP feed, a JASON connection or any other standard way of transmitting data). The data center itself will take care that access to the participating sites is hidden to the outside world but that the sites can still be queried as a single resource. In practice this will be difficult to do, as, for instance, the Volterra database is in MS Access. The default Access driver only supports local JDBC connections (via a so-called JDBC-ODBC bridge), as the connection URL just specifies the file path to the Access database .mdb file. For a client-server model, Access requires us to purchase a so-called type 3 or 4 Access JDBC driver that supports connection to Access from a remote client. Therefore, at least Access and OGSA-DAI would need to be run on the same host, or else we would need to break with our original assumption and transform the source database. This would mean migrating the Access data source into a non-desktop database like MySQL.

A good analogy for illustrating the benefit of this work is a map, where each dataset represents a small area—say, a few houses within a street. If you integrate a few of them it is of limited use, but after a certain point is reached you will have enough information to navigate your way from A to B. The data resources used in the project are just three examples—there are many small, scattered yet related data resources that would benefit researchers if they were linked along the lines described above to form a virtual data center for researchers, uniting scattered and inaccessible data resources and enabling them to ask questions that they would not have been able to ask otherwise. It has frequently been argued on both sides of the Atlantic that, in such cases, the whole has the potential to be much greater than the parts.[1] The utility of these datasets will increase greatly once a certain critical mass is reached.

[1]E.g. Keith W. Kintigh, "The Promise and Challenge of Archaeological Data Integration," *American Antiquity* 71 (3) (2006): 567–568.

6 RESULTS AND NEXT STEPS

LaQuAT was a collaboration between the OGSA-DAI development team at EPCC, the domain experts in classical archaeology in CCH, and experts in the area of information science and e-infrastructure at CeRch. Although the aim was to get these researchers to engage with the technology, it is good to note that not only was the project of benefit to the researchers, but the challenge of dealing with this material also led to improvements in OGSA-DAI, namely improved database drivers and enhanced XML functionality.

We have successfully implemented the kind of distributed queries that were originally envisaged. XML-database integration is still ongoing however. But we have learned a key lesson, which is extremely interesting both from a database design point of view, and from the point of view of linking inscriptions and archaeological data in ways indicated in the introduction. Union queries, which essentially combine different fields in different databases and allow searching as if they were a single field, are very useful. However, joining tables across different databases is likely to be far more relevant to the questions that archaeologists want to ask, due to the nature of the data that we are dealing with. An example of such a 'join query' could be 'Which laws, catalogued in database X, were enacted in the time period of such-and-such an official, whose dates are attested in database Y?'. Such multi-dimensional and nuanced queries are both more technically challenging that union queries, and likely to be of more interest to the archaeological community.

A major factor is that the information in these datasets is fuzzy and uncertain. For example, suppose that two separate databases, or indeed two rows within the same database, refer to a person named Licinius. Do they refer to the same person? At an intuitive level, this is a matter of judgment for the researcher, based on evidence both within the databases and external to them. Dates are represented in a variety of different ways, for example in relation to the reign of the emperor, or in terms of the two Roman consuls for a particular year, or in other forms that vary depending on the region. It is no easy matter to compare them or map them onto modern date terminology. Again, such decisions are subject to the interpretation of the individual researcher.

A particularly challenging issue being investigated is that of handling different levels of uncertainty in temporal data; some dates are extremely precise—even to the day—whereas others are very vague—perhaps to a span of 50 or 100 years. It is therefore not a simple matter of allowing researchers to query over an integrated view of the different databases. It is not useful simply to identify data columns in different databases and join them using a distributed query. In research terms, the results that are returned from one database may well influence the questions that are

asked of others—so a relatively straightforward query joining databases can, on examination, expand into a more complex workflow with the researcher at the center.

In the immediate term, the LAQUAT project is creating a demonstrator with a web interface so that researchers can investigate it for themselves. So far we have been hosting everything locally. However, we are investigating using the UK's National Grid Service (NGS) for hosting OGSA-DAI servers and some data sets remotely. We are aware, however, that this will raise significant security and IPR issues.

As well as databases, there is much other material in this sphere which is marked up in XML or other mark-up, so enhancing OGSA-DAI's support for XML is very important to us. This will also enhance linking the kinds of databases in this demonstrator with more archaeological material, much of which is supported by some kind of XML encoding. We would wish to include other datasets in this. As well as datasets relating specifically to inscriptions and papyri, there are many relevant archaeological databases, and other data resources that would allow better use to be made of the information such as the Lexicon of Greek Personal Names (www.lgpn.ox.ac.uk/), the American Numismatic Society's coin database (www.numismatics.org/ Collections), and prosopographic databases.

ACKNOWLEDGEMENTS

We would like to thank the ENGAGE initiative (http://engage.ac.uk/engage), which funded the LaQuAT project, and the Joint Information Systems Committee (JISC), which in turn funded ENGAGE. We would also like to thank OMII-UK (http://www.omii.ac.uk/), which develops and supports the OGSA-DAI software, and in particular its Director, Neil Chue Hong, for help, support and encouragement.

BIBLIOGRAPHY

Antonioletti, Mario, Tobias Blanke, Gabriel Bodard, Mark Hedges, Alastair Hume, Mike Jackson, and Shrija Rajbhandari. *Linking and Querying Ancient Texts (LaQuAT): Final Report*. Project report (forthcoming).

Lock, Gary. *Using Computers in Archaeology: Towards Virtual Pasts* (London: Routledge, 2003).

Grant, A., M. Antonioletti, A. C. Hume, A. Krause, B. Dobrzelecki, M. J. Jackson, M. Parsons, M. P. Atkinson, and E. Theocharopoulos. "OGSA-DAI: Middleware for Data Integration: Selected Applications", *IEEE Fourth International Conference on eScience, 7–12 Dec. 2008*: 343.

Kintigh, Keith W. "The Promise and Challenge of Archaeological Data Integration," *American Antiquity* 71(3) (2006): 567–578.

Moffett, Jonathan. "Computers in Archaeology: Approaches and Applications Past and Present," in *Computers for Archaeologists*, ed. Seamus Ross, Jonathan Moffett and James Henderson, 13–39. Oxford: Oxford University Committee for Archaeology Monograph No. 18, 1991.

Snow, Dean R., Mark Gahegan, C. Lee Giles, Kenneth G. Hirth, George R. Milner, Prasenjit Mitra and James Z. Wang. "Cybertools and Archaeology," *Science* 311 (2006): 958–959.