DHQ: Digital Humanities Quarterly

2009
Volume 3 Number 4

Grid-enabling Humanities Datasets

Mark Hedges <mark_dot_hedges_at_kcl_dot_ac_dot_uk>, Centre for e-Research, King’s College London

Abstract

The term “grid-enabling” is sometimes (or even often) used without a clear idea of what is meant. In this article we attempt to clarify some of the possible meanings of grid-enabling data resources. In particular, we examine how researchers in the humanities may benefit from using such approaches, and examine some concrete case studies in which grid technologies have been used to support data-driven research in the humanities.

Introduction: What Do We Mean by Grid-enabling Data?

Like many vogue expressions of technological origin, “grid-enabling” is a term that is sometimes (or even often) used without a clear idea of what is meant. While it may seem inappropriately techno-centric to take as point of departure a technology rather than the issues that one might address with the technology, the uses and abuses of this expression may justify this inversion as a way of clearing the ground of confusion. To begin with, then, let us consider for a moment what we mean when we talk about grid-enabling data resources, before we move on to the more specific topic of humanities data, and provide some concrete case studies. 1

Grid computing has been defined in a variety of ways, encompassing a variety of types of distributed computer systems. [1] The term *grid* alludes not to the image of a linked network of computers, but rather to the analogy of public utilities, for example an electricity grid, where a consumer can connect a diversity of electrical appliances, making use of open and standard interfaces (e.g. a plug), and consume electricity, without knowing or caring about its origin. Similarly, a diversity of electricity sources (coal, nuclear, solar) can contribute electricity to the grid, without reference to who is going to consume it. From the point of view of the consumer, this diverse reality is virtualised as a single source of electricity. A related but not identical concept is “utility” or “cloud” computing[2], used to describe services such as Amazon's EC2 and S3[3], which respectively provide computation and storage services. However, despite the similarity of the metaphor, the functionality that grids aim to provide is considerably more extensive. 2

Irrespective of these definitions or metaphors, when people speak of grids they frequently mean computer systems that have in common the fact that they are powered by one or more of a particular group of technologies, which fall under the category of “grid middleware.” In this context, to grid-enable data resources means to make them available via a grid, or by using some form of grid middleware. But this technical answer doesn’t help us much in understanding what we mean. Unlike technologists, researchers do not in general make the effort to use such technologies for their own sake. We need rather to look at the question of what we want to achieve by making resources available in such a fashion, and what researchers and other users will gain thereby. So what can we enable by grid-enabling data? One basic step is to enable sharing of a data resource, widening its availability and increasing its use and effectiveness. For instance, a data resource may be held on a departmental or institutional file system, or in the “deep web,” that is to say regions of the World Wide Web that for one reason or another are not accessible by search engines, and thus while in principle accessible are not discoverable using generally applied techniques. Resources of this kind can be shared by putting them on a publicly available website within the reach of web crawlers, or in an institutional repository, which can themselves be opened up to web indexing agents. Such access is a necessary precondition, though not in itself sufficient to constitute true grid-enablement. 3

More significant possibilities exist in expanding research capacity, through the virtualisation of data and access to data, by means of abstracted and standardised interfaces and protocols. This virtualisation may operate with respect to a number of different factors: 4

- Location. Access is provided independently of where the datasets reside.
- Autonomy. Data may be governed by independent management regimes, owned by different communities and subject to different rights. Access is made more uniform while respecting the integrity of the original data and the environments in which it is managed.
- Heterogeneity, both the infrastructural heterogeneity of the storage, and the structural heterogeneity of the data. Virtualisation means that datasets do not need to be accessed in possibly idiosyncratic ways.

Virtualisation can hide “irrelevant” (for whatever purpose we have in mind) differences between data resources, giving the user more seamless access to them. Distributed, autonomous and heterogeneous datasets can be federated and regarded as a single resource, enhancing the visibility of the data and multiplying the uses to which it can be put. 5

Grid Applications in Data-driven Research?

In order to see what grid-enabling may mean, and what it can do for the humanist, let us examine some applications of grid technologies in scientific disciplines, where they have been more widely used. 6

The earlier focus of grid technologies was a computational one, enabling the distribution of very large computational tasks and simulations in the “big” sciences, such as physics or the environmental sciences. The grid-enabling of data in these disciplines was concerned to a great extent with enabling fast access to and transfer of very large data sets distributed over multiple, collaborating research centres. 7

The capture of extremely large data sets and their availability via the Internet has also led to a new model of carrying out scientific research. Instead of the direct observation of natural phenomena (experimental science), the creation of (usually) mathematical models accounting for natural phenomena (theoretical science), or the generation of knowledge through large scale simulations (computational science), the researcher works on, gains knowledge from and adds value to already existing collections of digital data. We meet in particular the model of the *in silico* experiment (Goble et al. 2006), where data from multiple, distributed and heterogeneous sources is processed by software tools to create new information and generate new discoveries, for example, sky survey data held in “virtual observatories” [Walton et al. 2005], or databases of protein or genomic data.[4] 8

These examples are apposite for the humanities. On the one hand, increasingly numerous corpora of data are being produced by research projects in the humanities, as well as by digitisation programmes at libraries, archives and museums, producing digital surrogates of valuable primary source material for humanities research. On the other hand, however, the creation of digital data and the tools to process that data need to advance hand in hand. Although many data sets of value to the humanities have been produced, a situation in itself of value to researchers who (for example) no longer need to travel to a particular archive and search through documents by hand, the tools (whether grid-based or not) to carry out new modes of research using that data have lagged behind. 9

[Blanke et al. 2009a] contains a survey of recent and current projects in the UK that are applying grid technologies and other e-science techniques to humanities disciplines, and most of this work may be described as data-centric. It is noteworthy that, among these projects, those that involve a significant degree of intensive computation do so precisely because of the character of the data they are addressing[5], rather than because of a need for fast access or because of the size of the data sets involved (although the latter should not be overlooked[6]). Data-driven research in the humanities is, and will continue to be, motivated by a different set of requirements and use cases that arise from the very nature of the research data and of the research that is based on it. 10

Integrating Humanities Data: A Case Study

To illustrate the particular challenges raised by data in the humanities let us consider a specific example. The LaQuAT 11

(Linking and Querying Ancient Texts) project [Blanke et al. 2009b] investigated the use of grid middleware for providing integrated views across diverse humanities data resources. Specifically, these resources related to classical epigraphy and papyrology, and included relational databases with different schemas implemented using different data technologies, together with a corpus of XML data, all of them genuine resources developed by research projects in the humanities. These resources were used as exemplars because although they contain data which overlaps in terms of time period, geographical extent and area of historical interest, they were created by different research groups for different purposes. They use different vocabularies to describe the data they contain, they are structured and presented differently, they use different field headings and, in many cases, the same or similar objects have been entered using different epigraphic terms and/or metadata. In other words, they realistically illustrate the actual variability of humanities data sets. The resources chosen were:

- The Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV: http://www.rzuser.uni-heidelberg.de/~gv0/), a database containing metadata on some 55,000 papyri, mostly from Roman Egypt and its environs, including bibliography, dates and places (e.g. findspots and provenances). [7]
- Projet Volterra http://www.ucl.ac.uk/history2/volterra/, a database of Roman legal pronouncements and associated metadata, from various sources, whether epigraphic, papyrological, juristic or literary.[8]
- The Inscriptions of Aphrodisias http://www.insaph.kcl.ac.uk/index.html, an XML corpus of 1500 inscriptions from the ancient city in Asia Minor, including transcribed texts and metadata marked up using EpiDoc TEI, as well as images.[9]

These resources were just three examples from a larger pool of related datasets that might have been included in a larger scale project to support the classics researcher; they were selected in order to investigate the feasibility of the approach and to identify issues that might arise. Considering both these three datasets and the researcher’s broader data environment, we may make the following observations: 12

- Data formats are very diverse, and involve multiple media and standards. Databases rarely follow standard database schemas. The use of mark-up can vary significantly, particularly in resources developed before much effort had been made towards standards (such as EpiDoc), but natural variation occurs even in applying these standards.
- The material may be highly complex, with many structural and semantic relationships both internal (for example within a TEI document) and contextual. The interpretation of an object (e.g. an inscription) may depend on its relationships to other resources and collections (e.g. other inscriptions, literary texts, archaeological surveys, concordances), which are moreover not necessarily digital.
- Data may be incomplete, or indeed incompletable in cases where the capture of the data cannot be repeated nor the data enhanced to fill in the gaps. For example, an inscription may be damaged, the provenance of a papyrus not recorded, a corpus of texts fragmentary, the date of an event unknown.
- Data may be fuzzy or uncertain, or even contradictory. For example, there may be several sources for the date of an event, with various degrees of precision (to the year, to the decade) and various degrees of reliability.
- The resources are not easily available for use. In many cases, they are locked away on departmental machines; in other cases they are “published” on a website but not in a way that makes the resources particularly usable by a researcher.
- Even when a resource is available it is often available only 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.
- The 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. Consequently, any integration initiative must respect this autonomy and integrity, if it is to be successful.

We would not argue that such issues arise with respect to data only in the humanities, nor that all humanities data can 13 be characterised in this way. These issues will however be recognised by a significant number of researchers in humanities disciplines.

While the LaQuAT project demonstrated the feasibility of such an approach, we would not claim that integrating these three data resources in itself facilitated useful research in the classics. However, these resources were just three examples; there are many small, scattered yet related resources that would be much more useful to researchers if they were linked along these lines. Their utility would increase greatly once a certain critical mass is reached, and together they would form a whole much greater than the sum of the parts [Kintigh 2006], enabling researchers to ask questions that would not otherwise have been possible. An analogy might be a map, where each dataset represents a small area, say a few houses within a street; integrating a few of them is of limited utility, but after a certain point is reached there will be sufficient information to navigate from one place to another. Ultimately, this approach could lead to a broader vision of a "virtual data centre." 14

It is however difficult to address the generation of meaningful links between these data resources in a purely automated fashion. The fuzzy, uncertain and interpretative nature of the data complicates the semantics of integration, and makes it difficult to describe the semantics of the relationships between data sets. For example, it was not always clear whether similarly named columns in independent databases really represented the same sort of information and could validly be linked. In some cases there were deeper semantic issues, for example when two independent data sets contained contradictory information. While automated linking of data proved problematic and uncertain, it would be useful to describe, rationalise, or quantify this uncertainty. Often this is a matter of judgement for the researcher, based on other evidence both internal and external to the resource, and in such a situation researchers will want to define for themselves how resources relate to one another. The results of one query may, taken together with other information available to a researcher, influence the questions that are asked of others, and a more natural view would be a complex workflow with the researcher at the centre. 15

Digital Repositories and the Grid

One of the issues addressed in the previous case study was structural and semantic diversity. While there are a variety of standardisation activities with the aim of increasing interoperability between digital resources and enabling them to be used in combination, standardisation alone is unlikely to solve all problems related to linking up data. Humanists still have to deal with legacy data in diverse and often obsolete formats, and even when standards are used the sheer variety of data and research means that there is a great deal of flexibility in how the standards are applied. Moreover, standards are generally developed within particular disciplines or domains, 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. 16

Let us look now at technologies with the potential for managing data diversity. The outputs of research or digitisation programmes are increasingly being held in formally managed digital repositories, based on particular digital repository software, the most popular examples of which are EPrints, DSpace and Fedora.[10] In the UK, the Joint Information Systems committee (JISC) has funded several programmes to encourage the uptake and enhancement of such systems, and research into improving their utility. 17

In their earlier incarnations, repositories were used to manage relatively simple content, primarily research articles, sometimes less formal material such as presentations. However, digital repositories have been changing, both in the type of content that they hold, and the ways in which they are used. Repository software has become more sophisticated, allowing complex digital content to be stored in such a way that its internal structure and external context can be explicitly represented, managed, described and exposed. In particular, they are beginning to be used to manage research data in a variety of disciplines. This approach is useful for research in the humanities, and for the digitised resources, such as archives, that often form the primary sources of such research. It enables researchers, in collaboration with information scientists, to put a degree of formal order into our digital objects while not restricting their essential variation. 18

Digital repositories are likely to be a key component of any future e-infrastructure for the humanities, or indeed for any other discipline[11], not only providing a mechanism for making academic research available in away that does not require other institutions to acquire replicas of that work (e.g. journals), but also generic mechanisms for representing the vast body of potential material that traditional publishing does not accommodate easily. The topology of the emerging repository ecology is not as yet clear. On the one hand, in the UK at least there is a drive towards the creation of institutionally-based repositories[12] to hold the research outputs of that institution. However, if repositories are to manage complex material as indicated above, discipline-specific knowledge and expertise will be required, and we may expect that bodies specialising in certain fields may arise, along the lines of the now defunct Arts and Humanities Data Service in the UK.[13] 19

Whatever the outcome, given the increasingly cross-institutional and cross-disciplinary nature of much research, it is likely to involve material across repositories that are distributed and independently managed. The work undertaken by the grid community on the integration of structured information may be able to help us interact with these highly structured repositories and the complex data they contain, which will thus become virtualised data resources on a grid, like the storage devices and databases considered above. In particular such an approach may be used to hide the idiosyncratic heterogeneity of digital objects whose degree of standardisation will only decrease as their complexity grows, allowing data resources within repositories to be discovered, integrated and delivered in a form appropriate to the work to be done. 20

Another Case Study: gMan

The LaQuAT project concluded that there was a need for the human element in identifying connections between data resources. This led us to investigate the potential for a more interactive annotation environment that would allow researchers to create such connections and associate with them additional information, such as confidence levels or details of the evidence on which a decision was made. This provides another example of the use of grid technologies to facilitate the integration and re-use of humanities data, based on the promising gCube, which was developed within the European D4Science project[14], and which moreover illustrates the potential synergies between grid technologies and digital library systems. The gCube system is designed to support research by facilitating the creation of on-demand data-centric Virtual Research Environments (VREs) that are tailored to the needs of specific research groups, and are built on top of grid infrastructures that use the gLite middleware, the most notable example being the European EGEE infrastructure.[15] These environments provide virtual repositories that allow pre-existing data resources, diverse in terms of formats and metadata standards, to be combined, manipulated and annotated. The utility of the gCube system for the humanities is being investigated by the gMan project.[16] 21

To take a simple example from LaQuAT — the symbol *?* was frequently used in the original data sources to indicate that an entry in a database table was unknown. An annotator might fill in these unknown entries, and either determine their value fully or associate an uncertainty measure with the inserted data, as well as recording the basis on which the decision was made. Suppose, for example, that a researcher is looking for a date in the Volterra database that turns out to be unknown. In an annotation environment, the researchers might replace such an unknown entry with an actual entry of 100 BC, adding 80% as a confidence value, and providing links to the external data sources on which the decision was based. To take another example, it may be impossible to determine in automated fashion whether two occurrences of the same name, or two non-identical variants of a name, refer to the same historical person. The annotation environment would allow such connections to be expressed, tagged with additional information about confidence and provenance. 22

Our investigations led us to conclude that there is a need among at least some humanities researchers for tools supporting collaborative processes that involve access to and use of complex, diverse and geographically distributed data resources, including both automated processing and human manipulation, in environments where research groups (in the form of “virtual organisations”), research data and research outputs may all cross institutional boundaries and be subject to different, autonomous management regimes. These tools may form part of a collaborative Virtual Research Environment (VRE) (see [Fraser 2005]) and may focus on a specific discipline or even research question. Such 23 environments may, but need not, make use of grid middleware. Thus gCube provides a mechanism for creating VREs that exploits existing grid software and infrastructure.[17]

Conclusion

It is probably still not clear what people mean by grid-enabling data; indeed it is safe to say that they use it to mean different things, with varying degrees of precision, in different contexts. We have seen that the term has its origin in a certain class of technologies, and that researchers' focus when using grids in data management was at the outset a technical one, addressing fast access to very large, distributed datasets on virtualised, "joined up" storage devices. 24

Although both the gMan and LaQuAT experiments could be described as "grid-enabling" the data in question, as each makes use of a grid technology to link up the datasets, the aims and results of the two projects are quite different. Thus the grid should not be viewed as a single technology that joins things together, but rather as a range of different technologies, possessing among themselves a "family resemblance," which can be used to join things together in a range of different, but related, ways. Moreover, it is not the only sort of technology that can play such an integrating role; Web technologies are a more mainstream example, particularly the emerging Linked Data technologies.[18] [19] 25

Whatever combination of technologies is used, however, data in the humanities cannot be addressed in isolation from the researchers who create and use the data, and who also, of course, interpret the results. Research in the humanities can be highly interpretative, and will continue to be so even when supported by technological or scientific methods. It is necessary to link up not only data, but also services and researchers — in the plural. Research in the humanities need no longer be an activity carried out by a single scholar, but rather by collaborating researchers interacting within an extended network of data resources, digital repositories and libraries, tools and services, and other researchers, a shared environment that facilitates and sustains collaborative scholarly processes. 26

Of greatest interest to humanities researchers, whose data resources may be more notable for their complexity, diversity and fuzziness than for their size, is "grid enabling" in the sense of joining and modelling disparate information. Arguably this contrasts with the emphasis of the grid's origins in "big science," where the principal challenges involved processing queries and distributing very large datasets. By equipping humanities data to be interoperable using grid technologies, we can aggregate the answers it provides to research queries, and thus strengthen existing research practice and methods. In the examples discussed above, "grid-enabling" amounts to integrating or mediating between autonomous resources with disparities in structure, form or semantics, thereby enabling research driven by that broader corpus of data that would not otherwise have been possible. Through the creation of Virtual Research Environments, grid technologies can also be used to create that allow humanities researchers to manipulate and combine datasets from scattered sources. 27

Notes

[1] Useful definitions of the term *grid computing* are offered in [Foster 2002]; http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html; [Buyya & Venugopal 2005]; [Foster & Kesselman 1999].

[2] For more information on *utility computing* see [Yeo et al. 2007] and [Hand 2007].

[3] See http://aws.amazon.com/.

[4] See http://www.wwpdb.org/ and http://insdc.org/.

[5] Note in this connection the ReACH (Researching e-Science Analysis of Census Holdings) workshops held at UCL (see Section 2.2 of [Blanke et al. 2009a] and http://ahessc.ac.uk/files/active/0/ReACH-report.pdf), and the Medieval Warfare on the Grid: The Case of Manzikert project at the University of Birmingham (see Section 3.3 of [Blanke et al. 2009a] and http://www.iaa.bham.ac.uk/research/projects/manzikert/index.shtml).

[6] The volume of digital content available to the humanities and cultural heritage communities has increased rapidly in recent years. For example, much material is being produced by the large scale digitisation of archives or primary sources, particularly in relation to high resolution

images or audio-visual material; disciplines that use “scientific” techniques such as LIDAR and high-resolution 3D laser scanning, such as archaeology or the study of material culture, generate very large datasets (see http://ads.ahds.ac.uk/project/bigdata/final_report/bigdata_final_report_1.3.pdf); modern archives are increasingly “born digital,” forming rapidly expanding resources of unprecedented detail for future historical research.

[7] See http://www.rzuser.uni-heidelberg.de/~gv0/Texte/HGV-Texte.html.

[8] See http://www.ucl.ac.uk/history2/volterra/index.htm.

[9] See http://insaph.kcl.ac.uk/iaph2007/.

[10] See http://www.eprints.org/; http://www.dspace.org/; http://www.fedora-commons.org/

[11] See Tony Hey’s keynote address “e-Science and Scholarly Communication” at Open Repositories 2007 (http://openrepositories.org/2007/program/speakers).

[12] See http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue.aspx.

[13] See http://www.ahds.ac.uk/.

[14] See http://www.gcube-system.org/.

[15] See http://www.eu-egee.org/; http://glite.web.cern.ch/glite/.

[16] See http://gman.cerch.kcl.ac.uk/.

[17] Another example from the humanities is the TextGrid project, which has created an environment for the collaborative editing, annotation, analysis and publication of texts. See http://www.textgrid.de/.

[18] See http://www.semanticgrid.org/presentations/DEROURERepo3.pptx, in particular Slides 28-31.

[19] See http://linkeddata.org/.

Works Cited

Blanke et al. 2009a Blanke, T., M. Hedges and S. Dunn. “Arts and Humanities e-Science--Current Practices and Future Challenges”. *Future Generation Computer Systems* 25: 4 (2009), pp. 474-480.

Blanke et al. 2009b Blanke, Tobias, Gabriel Bodard, Stuart Dunn, Mark Hedges and Shrija Rajbhandari. “LaQuAT: Integrating and querying diverse digital resources in classical epigraphy”. Presented at *CAA 2009. Proceedings of Computer Applications and Quantitative Methods in Archaeology* (2009). http://laquat.cerch.kcl.ac.uk/.

Buyya & Venugopal 2005 Buyya, Rajkumar, and Srikumar Venugopal. “A Gentle Introduction to Grid Computing and Technologies”. *CSI Communications* 29: 1 (2005), pp. 9-19. http://www.buyya.com/papers/GridIntro-CSI2005.pdf.

Foster & Kesselman 1999 Foster, Ian, and Carl Kesselman. *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco: Morgan Kaufmann Publishers Inc, 1999.

Foster 2002 Foster, Ian. “What is the Grid? A Three Point Checklist.” July 20, 2002. http://www.mcs.anl.gov/~itf/Articles/WhatIsTheGrid.pdf.

Fraser 2005 Fraser, Michael. “Virtual Research Environments: Overview and Activity”. *Ariadne* 44 (July 2005). http://www.ariadne.ac.uk/issue44/fraser/.

Goble et al. 2006 Goble, Carole, Katy Wolstencroft, Antoon Goderis, Duncan Hull, Pinar Alper, Philip Lord, Daniele Turi, Robert Stevens, Jun Zhao, Khalid Belhajjame and David De Roure. “Knowledge Discovery for in silico Experiments with Taverna”. In Christopher Baker and Kei-Hoi Cheung, eds., *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Springer, 2006.

Hand 2007 Hand, Eric. “Head in the Clouds”. *Nature* 449: 963 (2007). http://www.nature.com/news/2007/071024/full/449963a.html.

Kintigh 2006 Kintigh, Keith W. “The Promise and Challenge of Archaeological Data Integration”. *American Antiquity* 71: 3 (2006), pp. 567-568.

Walton et al. 2005 Walton, N. A., A. M. S. Richards, P. Padovani and M. G. Allen. “The Virtual Observatories: a major new facility for astronomy.” *Proceedings of the International Astronomical Union* 1: 398-403 (2005).

Yeo et al. 2007 Yeo, Chee Shin, Buyya Rajkumar, Marcos Dias de Assunçao, Anthony Sulistio, Jia Yu, Srikumar Venugopal and Martin Placek. “Utility Computing on Global Grids”. In Hossein Bidgoli, *The Handbook of Computer Networks*. New York: John Wiley & Sons, 2007. http://www.gridbus.org/papers/HandbookCN_Utility_Grids.pdf.