

Report: E-Science Scoping Seminar – History

Prepared for the Arts and Humanities Data Service
by:

Mark Greengrass
Humanities Research Institute
University of Sheffield



CONTENTS

ACKNOWLEDGEMENTS

1. INTRODUCTION	3
2. DEFINITIONS AND AGENDA	3
3. THE HISTORICAL DISCIPLINES AND E-SCIENCE	4
4. E-SCIENCE TOOLS FOR HISTORIANS	8
5. E-SCIENCE TECHNOLOGIES AND HISTORIANS	9
6. E-SCIENCE ENVIRONMENTS AND THE HISTORICAL DISCIPLINES	10
7. CONCLUSIONS	12

APPENDICES

Acknowledgements

The E-Science Scoping Seminar in History was administered by the Arts and Humanities Data Service Team, in particular the Project Research Assistant, Luke Blaxill, and the Administrative and Events Assistant, Katrin Weidemann. In Sheffield, the staff of the Humanities Research Institute, in particular its administrator Julie Banham, contributed greatly to its success. Presentations were given at the Seminar by Simon Hodson, Jamie McClaughlin, Luke Blaxill, Sheila Anderson and Dolores Iorizzo. Their assistance is gratefully acknowledged here.

1. Introduction

1.1 The seminar took place in the Humanities Research Institute in Sheffield on Wednesday 14 June.

1.2 Those present were:-

Sheila Anderson – Director, AHDS, King’s College, London

Luke Blaxill – Project Research Assistant, AHDS, King’s College, London

Louise Craven – Programme Manager, Access to Archives (A2A), Records Management and Cataloguing Division, The National Archives

Matthew Davies – Director, Centre for Metropolitan History, Institute for Historical Research, University of London

Paul Ell – Director, Centre for Data Digitisation and Analysis, Queen’s University, Belfast

Stuart Dunn – Research Associate in E-Science Methods, Arts and Humanities E-Science Support Centre [EHSSC], King’s College, London

Mark Greengrass – Associate Director, AHRC ICT Methods Network, University of Sheffield

Simon Hodson – Co-Director, VRE in Political Thought, University of Hull

Lorna Hughes – Manager, AHRC ICT Methods Network, King’s College, London

Dolores Iorizzo – Manager of Arts and Humanities, The London E-Science Centre

Jamie McLaughlin, Technical Officer, Humanities Research Institute, University of Sheffield

Katrin Weidemann, Administrative and Events Assistant, AHDS, King’s College, London

Matthew Wollard – Director, AHDS History, University of Essex

John Young – Research Officer, The Newton Papers Project

Contact details are provided in **Appendix One**.

2. Definitions and Agenda

2.1 **Definition of E-Science.** Sheila Anderson furnished a broader working definition of e-science than that offered by the National E-Science Centre [‘In the future, e-Science will refer to the **large-scale** service that will increasingly be carried out through **global collaborations** enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large **data collections**, very large-scale **computing resources** and high-performance **visualization** back to the individual user scientist’ <http://www.nesc.ac.uk>]. She wanted the seminar to concentrate on E-Science as

an **infrastructure** that involves **tools, technologies, networks** and **methods** to support research and learning.

- 2.2 **The Agenda of the Seminar.** Sheila Anderson outlined the purpose of the seminar: to inform the call for larger research-grants in E-Science and 6 research studentships from the EPSRC [Engineering and Physical Sciences Research Council], which is managing the E-Science programme overall, that will emerge in August 2006 of the possibilities and needs in the various Arts and Humanities disciplines. These needs might well involve e-infrastructure development, virtual research environment growth, specific tools, etc. They may grow out of a particular research programme, or be independent of it. The studentships may be self-standing, or linked to a wider programme.

The aims of the seminar were:

- a way of raising awareness and understanding about the potential for E-Science in the Arts and Humanities.
- a route by which significant research funding can be placed at the disposal of Arts and Humanities research.
- a horizon-raising exercise, encouraging the Arts and Humanities disciplines.

The outcomes of the seminar were:

- to identify where E-Science challenges and potential are likely to be most fruitful in the History domain.
- to determine immediate priorities for Historians.
- to delineate who the other contacts with potential E-Science interests might be.

3. **The Historical Disciplines and the Challenges of E-Science**

- 3.1 **Disciplinary Background and the Challenges of E-Science.** Mark Greengrass outlined the ‘challenges’ of E-Science to History in terms of its disciplinary background [see **Appendix Two**]. The application of E-Science to History was in its infancy, still at the stage of ‘research programme conceptualization’, and not yet at the stage of ‘pilot research project development’ let alone the consequential impact of those upon the culture of the discipline. The discussion around this paper focused on issues of E-Science as a means of:-
- connecting historical data (‘large’; and ‘small’)
 - connecting historians
 - connecting historians and data
- 3.2 **Connecting historical data: large datasets.** E-Science has, as developed in the hard-sciences, been a means for interrogating very large scientific datasets in highly sophisticated query modes:-
- The application to **quantitative** historical data is not likely to be the largest application of E-Science for historians in the next five years. One

of the largest datasets currently available from AHDS History – the 1881 Census – is 30 million records. Even accessing it remotely on a laptop, a fairly complex SPSS query recently only took about 17 minutes to perform [Woollard]. If we take into account the likely increases in local storage capacity and even a modest increase in processing capacity, all but the most complex of interrogations of quantitative data are likely to be handled by local micro-processing capacity. Even the current project to compare the census data from seven separate censuses (dir: Melissa Terras), one which involves a matrix of c.30 million variables, can be feasibly handled by a powerful server with adequate memory at today's technological specifications. In addition, the metadata which encapsulates a lot of the big social-science type material currently available is never likely to be compatible enough to be able to analyse 'all in one go', so that sequential and multiple processing of parts of the data is likely to remain integral to its interrogation. The real challenge for E-Science history in respect of large datasets is **availability**. The data tends not to be online. The example was cited of the 30 million published Irish census statistics [Ell]. For Irish historians, they are not readily consultable. Although available from the AHDS, what is needed is an E-Science **infrastructure** to make such materials straightforwardly usable by historians.

- The application to **textual** historical data is more complex. Historians are dependant on different suppliers for their textual data in electronic media (commercial producers: government bodies: libraries: archives: private sources, etc). They are rarely consultable in one place and at one time. The problem of distributed textual historical data is growing exponentially with the availability of material in electronic form. Very large quantities of text, and (in particular) large quantities of distributive historical text, are likely to require the application of E-Science technologies and methods in the near to medium-term future. This has challenges for the development of appropriate **tools** and **technologies** since there are complex metadata compatibility issues as well as important user-demands to bear in mind.
- The application to **visual** historical data is also complex. Historians use visual data in different ways. Visual data can sometimes simply be (and may well, in the future, increasingly) imaged files of original documents, often with little by way of metadata content. It can also, however, be in the form of images (photographic; audiophonic; cinematic: videographic) from the more recent past. For cultural historians, digitized images of cultural objects are also of great value. These are all likely to be significant growth areas in the discipline over the short to medium-term. But the very large size of some historical visual data (e.g. the Jean Froissart Project produces TIFF files of c. 133Mb for each image) is already posing a problem for consultation in an internet environment. For research on distributed image datasets, historians have a strong interest in pursuing the possibilities of E-Science technologies and methods. Here, too, there are challenges for the development of appropriate **tools** and

technologies since the metadata issues are here at a level of developing the appropriate metadata for E-Science applications, the user-demands have yet to be fully comprehended.

- The application to **mixed** historical data should be emphasized. Historians do not work on one sort of data to the exclusion of other sorts. Their domain is determined by period and problem, rather than source. Projects with ‘multimedia formats’ [Iorizzo] are likely to be of increasing significance, especially in an **international** context, where materials from libraries, archives, museums, and research practitioners becomes available in E-Research environments. Dolores Iorizzo prefigured one of these in a ‘History of Optics’ multimedia VRE, highlighted at the seminar, and Simon Hodson mentioned another, based on an international research project in the history of early-modern political thought.

3.3 **Connecting Historical Data: Small Datasets** There are very few, if any, research-based historians in the UK who do not use electronic media as part of their research desktop now [Young]. It is impossible not do so as a teacher in the discipline at UG and PG levels.

- But the engagement is with smaller datasets, often designed on the basis of relatively unsophisticated IT. One such resource is the RHS Online Bibliography of British and Irish History [<http://www.rhs.ac.uk/bibl/bibwel.asp>]. It contains over 407,000 records and is used on a daily basis by historians of British and Irish history. Although the server has recently been upgraded to provide Z39.50 (ISO23950) connectivity to conform to standard internet search query protocols and allow Endnote downloads, it does not provide automatic linkage to online versions of articles in the database. The technology exists for such a linkage. It is a matter of applying ‘best-practice’ E-Science protocols to the historical domain. The fact that it is not ground-breaking E-Science informatics should not stand in the way of an investment in their adoption.
- Many of the fundamental reference works for historical analysis remain unavailable in electronic media. For historians of Irish history, for example, the 23-volume bibliography of works up to the 1970s is only currently available in print in a limited number of locations [EII]. For historians working in the history of regions beyond the English language, the majority of the reference works for historical analysis are not available in electronic media – and a significant proportion (c.30%?) of the research practitioners in the UK work on non-English language historical domains. The E-Science agenda is only realizable in those areas of the discipline where the volume and depth of electronic media penetration is sufficient to warrant its application.

3.4 **Connecting Historians.** The advantages of the Access Grid and its personalized equivalent [the Personal Interface to the Grid; PIG] as an E-Science multilateral audio-visual platform are as evident in the historical discipline as elsewhere. The

majority of History departments in the UK are medium-sized by national standards and widely dispersed in HEI [see **Appendix Three**]. So the benefits of better lateral communication are evident. But the infrastructural investment in Access Grid suites of an appropriate size and scale, proximate to Humanities departments, still needs to be made in most HEI – not, of itself, a discipline-specific challenge. The technician costs for managing such suites are not insignificant, although PIG technology is now quite robust and user-friendly [Hodson]. The potential for making use of the distributed, often internationally recognized UK ‘centres of excellence’ [e.g. Centre for Metropolitan History; Institute for Historical Research, etc] in the discipline is, however, evident. But **those centres need to be identified** and equipped to take a lead in the E-Science area.

- 3.5 **Connecting Historians and Data** The potential for linking data and historians in an E-Science or E-Research virtual environment is considerable. The Seminar heard a first-hand report from Simon Hodson on the experience of the JISC-funded VRE programme in the History of Political Discourse, 1500-1800 [http://www.uea.ac.uk/his/research/projects/vre/ma/manchester_presentation.pdf] The project has experimented with the application of E-Science to PGT learning, linking teaching groups in Hull and East Anglia with texts, readings and an associated E-Learning environment delivered through an open-source web-based collaborative environment [SAKAI]. The advantages in teaching terms were highlighted: the ability to ‘not just display documents, it’s being able to tap into the computer to skim through them, to annotate them, highlight (passages), synchronously, as the discussion in the seminar goes on. We found that a very useful teaching application and research tool’ [Hodson]. It required **considerable preparation** of teaching materials, and a **specific adaptation** of pedagogic methods to the new environment. Two developments., for which British Academy funding has been successfully achieved, are a VReading Group and an eText Group. The first looks at an important **methodological issue:-** viz, how successfully an E-Science environment can be used in a collaborative research environment to explore a particular research question. The second examines a **methodological and tools issue:-** viz, how one might develop and utilize a tool for the group annotation of texts through wiki-based technology to understand/edit a particular historical text for research purposes. Both developments challenge established notions in the discipline of ‘authorial claim’ and ‘textual authority’. Although the RAE panel 62 is explicit about its willingness to accept electronic publications as output of the same weight as other forms of output, it is against the nature of the exercise itself to blur authorial claims. And there was a debate in the seminar about whether it was necessarily to the benefit of the discipline to have ‘textual authority’ challenged. The positive advantages of an environment in which ‘raw material’ could be ‘opened out’ or ‘democratized’ to a variety of interpretations in a ‘community forum’ were stressed [Iorizzo et al]. So too were the doubts as to whether there would not be considerable practitioner resistance, and whether ‘destabilizing’ the science represented by the ‘critical edition’ was in the research interests of the discipline

[Greengrass et al]. The issues of trust and authentication are ones that have already been faced in the physical sciences programme. Their lessons are ones that E-Science pilot projects in history could usefully take on board. The issues, in short, are not insuperable. But they lie at the heart of the **cultural challenge** of E-Science to established research practices in the historical domain.

- 4 E-Science Tools for Historians** In the course of our discussions, we highlighted several key areas where E-Science ‘tools’ development was likely to be significant for historians.
- 4.1 Historical Thesaurii and gazetteers.** Historians deal with fuzzy and imprecise data, which changes over time. Historical orthographies are unstable. The temptation to apply contemporary gazetteers and thesauri to historical materials is strong, and always misplaced. For cross-searching historical materials beyond their most basic interrogation, there needs to be
- a greater investment in historical thesauri and gazetteers. The example cited in the seminar was that of Parliamentary constituencies, where there is no standard list of the c.635 constituency names that can be applied in historic time. Yet it would be a ‘word of about a week’ [Woollard] to compile such a list.
 - a more effective network for publishing and distributing those that already exist. This includes making available the thesauri and gazetteers from other disciplines, e.g. archaeology [Davies]. There was general agreement that a lot of potentially useful name authority files were currently ‘locked up’ in personal research endeavours which could be profitably used in an E-Science environment to release value from distributed data.
- 4.2 Visualisation Tools.** The tools for the study of visual materials in digitized form are somewhat crude. E-Science affords the possibility for improving the comparison of one image to another, the searching and browsing of a digital library, and the collective annotation of a digital image. These would all be of potentially considerable use to cultural historians.
- 4.3 Text-Editing Tools.** The development of advanced E-Science text-editing tools is one that is both a ‘tool’ and also a ‘technology’. The adaptation of a wiki environment for group-text editing (see above, **3.5**) is a good example of a creative re-use of an E-Science tool already in existence. But there is also the more ambitious example of a portfolio of Grid-based text-editing tools conceived by the Arts and Humanities component of the German ‘D-Grid’ initiative, launched in February 2006. ‘TextGrid’ is a cooperative project to develop a modular platform for the study of texts [<http://www.d-grid.de/index.php?id=167&L=1>]. The platform will include many literary, but also some historical texts in its first phase. The modules include metadata standardization, collation, lemmatization, word-sorting and highlighting software packages for E-Science use. Such a development would, however, be only conceivable on a scale beyond that of the historical disciplines.

- 5 E-Science Technologies and Historians** The technologies particularly associated with E-Science and relevant to the historical disciplines are those involved with knowledge extraction and the ‘semantic web’. These technologies have attracted substantial informatics investment. Almost none of it has been deployed as yet to the benefit of the historical disciplines. The reasons for that are complex, and not simply the consequence of research strategies based on commercial logic. They have also to do with the nature of the historical discipline, and the fact that it functions with ‘essentially contested’ concepts that frustrate any simple application of knowledge-mappings to the field. Most accepted historical terminologies (‘feudal’; ‘renaissance’; ‘aristocracy’ etc) have a definitional fluidity, capable of making the term applied differentially, depending upon their period and location.
- 5.1 Ontologies and historical disciplines.** There is a knowledge-gathering exercise still to be undertaken on the development of ontologies and their application in the historical area. There are some ontologies that have developed for use in museums, libraries and the archives world that may have potential historical use in an E-Science environment. These include CIDOC-CRM [Conceptual Reference Model], where a SIG has been established for application in the cultural heritage area [Sinclair, Lewis and Martines, University of Southampton: <http://www2006.org/programme/files/xhtml/p174/poster174.html>] and the European Union’s Vicodi Ontology Project.¹ Whether the knowledge extracted by such ‘large-scale’ discipline-based ontologies is capable to answering the leading research questions asked by historians, however, remains an open question. It is easy to frame an E-Science environment that provides answers to the questions we already can answer. It is more difficult to frame one that answers the questions we cannot yet answer. That is because, in the historical disciplines, the next-generation questions are still generally asked by the individual scholar, framing the agenda in terms of the mirror-image to what we already know. The imaginary example [furnished by Greengrass] was of the hypothetical scholar, Dr Black, whose research agenda was the history of ‘darkness’ in the early-modern period. We know a great deal about ‘light’ – through the historians of science and others. It would be relatively straightforward to frame an E-Science agenda around the history of light. But it would be more challenging to do so around the ‘history of darkness’. By definition, Dr Black is a pioneer. Pioneers do not have ready-made datasets and collaborators.
- 5.2 Data Extraction and distributed data sets.** The limitations of applying meta-disciplinary ontologies in the historical domain may not, however, be applicable to their application in more locally-based and subject-specific areas. The seminar heard a report from Jamie McClaughlin on the ongoing work in the Armadillo

¹ . Reported in, *inter alia*, Gábor Nagypál, Richard Deswarte and Jan Oosthoek, ‘Applying the Semantic Web: The VICODI Experience in Creating Visual Contextualization for History’, *Literary and Linguistic Computing*, 20.3 (2005), 327-49.

History Data-Mining Project [<http://www.hrionline.ac.uk/armadillo/>]. The project uses a set of 12 online resources in eighteenth-century British social history to evaluate the benefits of automated data-mining techniques. It then explores the potential for applying these techniques to a wider collection of Arts and Humanities resources. The ontologies are based on name-authority files and gazeteers relevant to the material in question. Metadata from one set of materials is used to extract knowledge from other sets of materials which have not been marked up. Use is made of information redundancy to create probable relationships within the data. The data-extractors are standard Top-Bottom, Right-Left wrappers, but their development requires close collaboration with computer scientists. These techniques are particularly appropriate to E-Science application because the size of the files storing the 'relationships' contained in the ontology rapidly become very large indeed, and thus potentially most suitably utilized in a distributive environment across the Grid. It is also an obvious environment where the potential inter-disciplinarity of E-Science is evident. There are questions about the training of historical research practitioners in the use of such an environment, and also about their understanding the provisionality of the results to a particular query (that being the case for all machine-developed knowledge extraction). Data miners do not necessarily read historical documents in the same way as historians. Word frequency is no guide to the significance attributable to a document's historical meaning.

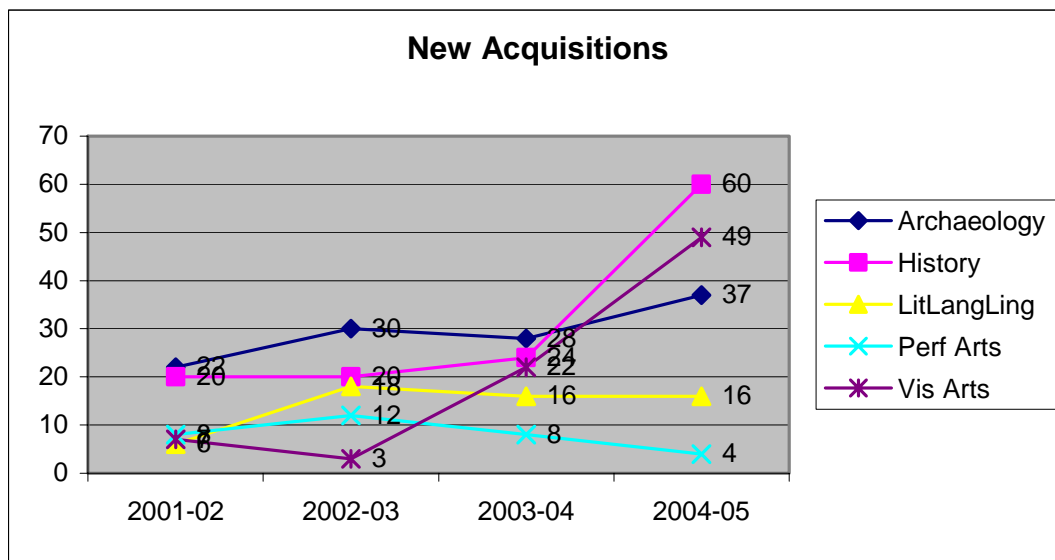
- 5.3 **European Framework to E-Science and History.** The importance of the European framework (and its Framework programmes) to the development of E-Science is evident. This will have an impact on the historical discipline in a number of different contexts:
- through the European integration of libraries, museums and archives
 - through European integration of multimedia technologies and its impact in the cultural heritage area
 - through European development of e-infrastructure.
- The footprint of the US Perseus Project [<http://www.perseus.tufts.edu/>] is an example of what a large-scale VRE can achieve, with over 2m hits per day [Iorizzo] testifying to its impact on learning and teaching. Such a project could be envisaged on an extra-national scale, bringing together European and international developments in these three areas.
6. **E-Science Environments and the Historical Disciplines.** The lesson from other innovative technologies is that a well-found demonstrator project/s has a motor effect on the wider community. They become exemplary of the potential of the technology, and its application in a particular discipline. The seminar heard of the positive and dynamic impact of the E-Bank of images in chemical crystallography as a model of what we should aim for. The seminar agreed that such an E-Science demonstrator environment would need to fulfill all, or some, of the following criteria:
- An area of the domain with substantial (preferably multimedia) datasets, already available in a variety of locations.

- An active research agenda, in which the questions being asked cannot readily be answered by traditional (individual) scholarly enquiry.
- An identifiable research community, preferably international, interested in investing time/energy in developing an E-Science environment.
- An environment in which a ‘vortex’ effect can be encouraged from the collaborative enquiries undertaken

61. **Virtual E-Science Environments and E-Science Desktops in History.** The innovative nature of E-Science should not be exaggerated. Many historians already regularly use a wide variety of electronic materials in their research and teaching. The problem is that they have to use them ‘serially’, with severe limitations to the potential for comparative searching, browsing, and little to no possibilities for more advanced manipulations of the electronic data. The creation of an E-Science ‘desktop’ that enabled the research practitioner in history to overcome these problems should be a longer-term objective. But it has to be developed from within the user-community and not, like e-learning programmes from learning technologists and administrators, whose vision of how learning happens in a particular discipline is often stereotyped.

6.2 **E-Science Environments and the Wider Historical Community.** The research practitioner in history does not function in a hermetically-sealed environment of professional historians. Family historians researching their genealogy, UG historians undertaking their dissertations and the wider community are part of the world of the historical disciplines, one where the TNA is aware (through A2A) of their importance [Craven]. There is a danger that the creation of an E-Science ‘lagoon’ of historical practitioners could become separated from this wider historical community, the latter being unaware of the interests and engagements of the former. History remains a ‘public’ discipline. The Research Councils have a stated mission to serve that wider community. Some advanced historical materials have been made available in electronic form for the wider community (e.g. the ‘Vision of Britain’ historical mapping Project – <http://www.visionofbritain.org.uk/maps/index.jsp>). The project is an example, however, of the tension between materials made available for that wider community, where data is presented in certain ways, with hard links implicit within it that are difficult to unpack and evaluate, and the development of E-Science research materials, where the interpretative judgments remain more open to scrutiny.

6.3 **E-Science, Reusability and Sustainability.** The potential significance of the E-Science agenda for repurposing and reusing electronically-created historical data was emphasized [Hughes]. There was a risk that Research Councils, needing to provide cost-benefit analyses to the Treasury, would regard the investments in resource creation and enhancement as difficult to justify. The public investment in electronic resource creation had been significant in the last five years – the rising tide of deposits at AHDS are testimony to it, with AHDS History among the largest:



Source; AHDS Annual Report, 2005

The challenge was to show how these resources might be used within an E-Science environment. At the same time, there was an important issue about the documenting of electronic resources for their future use in E-Science environments. The question of documenting work-flow in E-Science environments was equally significant.

7. **Conclusion: Exemplary History Developments for E-Science Applications.**

The Seminar provided, by way of conclusion a series of projects, tools, methodologies and environments of E-Science applications in History. They vary in scale and ambition. But they were all realistic and achievable. They should be taken as 'exemplary' of the considerable potential of E-Science in the Historical Disciplines.

- A tool for algorithmic searches of linguistic variations to identify place-name variations, and orthographic variations of words, spellings and thesauri. It might be a historically-specific form of the SKOS W3 Core guidelines, published in 2005 [<http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/>].
- A GeoStation data browser to provide an integrated mapping and interpretation environment for historians, of the kind developed for commercial usage in e.g. petroleum exploration [e.g. <http://www.auto-trol.ca/>]
- A historical place-name thesaurus for Britain, in which the place-name is chronologically tagged and becomes the link to the parish, sub-parish, field data in historic time.

- An E-Science multimedia environment for materials relating to the History of London, to include 3D reconstructions of the London historic built environment, archaeological evidence, historical data, etc.
- An E-Science version of the Electronic Cultural Atlas initiative [<http://www.ecai.org/>] in which, for example, it would be possible to link political, institutional and cultural materials to particular locations (e.g. the complex polity of Germany prior to its unification).
- Bibliographic linkages. Electronic bibliographies are currently available in a variety of locations and not cross-searchable, nor linked to the online resources that they catalogue.
- Metadata Standards. The E-Science agenda rests on the further development of metadata standards, especially for digital objects beyond text. This is important in the historical domain, and there are historical specificities to the attributions of meaning to cultural objects that need to be represented within metadata standards yet to be developed.
- Data Linkage and Project Reusability. Define 5-6 publically-funded electronic resources in the historical environment and link them in an E-Science environment and measure the consequential impact.
- A User-Friendly VRE toolkit, with an emphasis on text-based applications.
- To develop further the National Archives 'Global Search' environment for cross-catalogue searching [<http://www.nationalarchives.gov.uk/news/stories/115.htm>].
- Training Programme for E-Science in History to raise the skills, education and awareness base
- A demonstrator 'VRF' [Virtual Research Framework] in History – i.e. a cross between a 'VRO' [Virtual Research Organization] and a VRE [Virtual Research Environment] in which a body of scholars had access to 'tools, data and research collaborators'. One candidate for this VRF was in the early-modern History of Science, with its strong research community, the interdisciplinary nature of its materials and agenda, and the potential for developing tools for text-mining, concept-mapping, language technology applications, word-clustering, stylistic analysis, multi-lingual morphological tools, etc
- An E-Science-based registry of research practitioners in history in the UK, automatically populated from locally-stored CVs in a common format.
- An E-Science application to the variety of materials on the History of Parliament, where history, public administration, government and politics intersect. This is an area of central significance to the history and public life of the nation, where there is a proliferation of digital materials, stored distributively and with varieties of metadata, currently incapable of being put to effective use.

Appendix One

Contact Details of Those Participating in the Seminar

Sheila Anderson

sheila.anderson@ahds.ac.uk

Luke Blaxill

luke.blaxill@ahds.ac.uk

Louise Craven

The National Archives,
Kew,
Richmond,
Surrey, TW9 4DU

louise.craven@nationalarchives.gov.uk

Matthew Davies

Centre for Metropolitan History,
Institute for Historical Research,
University of London,
London WC1E 7HU

matthew.davies@sas.ac.uk

Paul Ell

Centre for Data Digitisation and
Analysis,
Queen's University,
Belfast

Northern Ireland, BT7 1NN

p.ell@qub.ac.uk

Stuart Dunn

Arts and Humanities E-Science Support
Centre
King's College,
London

Stuart.Dunn@kcl.ac.uk

Mark Greengrass

Department of History,
University of Sheffield
Sheffield S10 2TN

m.greengrass@sheffield.ac.uk

Simon Hodson

Department of History,
University of Hull
Hull,
East Riding, Yorkshire HU6 7RX
s.d.hodson@hull.ac.uk

Lorna Hughes

King's College,
Kay House,
7 Arundel Street,
London, WC2R 3DX.
Lorna.Hughes@kcl.ac.uk

Dolores Iorizzo

The London e-Science Centre
Imperial College
London SW7 2AZ
d.iorizzo@ic.ac.uk

Jamie McLaughlin,

Humanities Research Institute,
University of Sheffield
Sheffield S10 2TN

j.mclaughlin@sheffield.ac.uk

Katrin Weidemann,

katrin.weidemann@ahds.ac.uk

Matthew Wollard

AHDS History,
University of Essex
Colchester,
Essex CO4 3SQ

matthew@essex.ac.uk

John Young,

The Newton Papers Project,
Imperial College,
London SW7 2AZ

j.young@ic.ac.uk

Appendix Two

E-science Challenges in the World of Historical Studies

This document has been prepared to facilitate discussion at the e-Science Scoping Study Expert Seminar for History. It is written by a research practitioner in early-modern history, and based on my own experiences and viewpoints. It cannot claim to be representative of all aspects of these disciplines. Nor does it aim to be comprehensive. Feel free to disagree with all or any of what it contains. It is simply offered as a delineation of some of the key issues on which we shall need to focus in our expert seminar. I would like our discussions to be broad-ranging. But I would also like us to emerge with some practical recommendations and ways forward. The seminar offers us an opportune moment and a unique occasion to look at how best we can use E-science to our advantage in our particular discipline. But, in order to do so, we need to eliminate those, potentially alluring ‘yellow-brick roads’ that will not help us realize our individual and collective research objectives. I start with trying to encapsulate some of the distinctive features in the way in which historical studies are conducted. Then I examine the challenges and opportunities presented by ICT, outlining the areas where it has been of particular importance to us. Finally, I offer a series of headings for us to discuss the e-Science agenda, suggesting where we may most profitably take it forward in our own area.

1. Understanding the World of Historical Studies

There is not much which is unique to the methods and approaches of the historical disciplines. They are, to varying extents, shared with arts and humanities more broadly – and not least since every arts and humanities discipline has its historical research practitioners. But the nature of historical documentation leads to some relevant distinctiveness. Here are 10 points for starters:-

- there is no such thing as a ‘canon’ of historical research materials. There is no primacy of period or geographical place. Historians are necessarily interested in comparative dimensions, both temporal and spatial
- there is no innate primacy of one kind of material about the past, textual or non-textual, over any other kind of material. Just as the past is an exponentially expanding universe, so is its material
- all historical research conclusions are essentially contingent. In historical studies, there is no ‘last word’
- there is no accepted historical ontology for historical studies
- the nature of the historical research process involves sustained engagement with the records of the past. ‘Historicity’ is achieved through successive iterations of critical interrogation of the evidence
- because (in part) there is no ‘canon’ and no accepted ‘ontology’, collaborative historical enquiry has a relatively small basis on which to grow. However (especially where historical documentation requires specialist skills for its collection/collation/interpretation), collaborative work has proved very successful

- the ‘natural partners’ of historians are archivists, librarians, museum-keepers, holders of historical record of one sort or another. These are typically not in academic institutions. They are often in institutions which are relatively poorly endowed, and sometimes poorly equipped for ICT developments. So, along with research fabric of the AHDS, TNA, BL and IHR, there are numerous smaller repositories on which historians regularly rely for their raw materials. Much historical raw material was made available in published textual editions that have been in the public domain for over a century, and whose viability remains generally solid
- this does not preclude profitable engagement with other academic disciplines, both within the arts and humanities and beyond. But these are often ‘conceptual’ engagements, in which each participant seeks to learn something from the approach of the other, rather than being necessarily involved in research ‘collaboration’ as such
- historical research is not undertaken purely by academic historians. There is an important penumbra of research activity in particular areas undertaken by interested amateurs, on whose work historians often rely for some of their conclusions
- historical research has an important public dimension. Historians are stewards of the past, and that stewardship implies a responsibility for the way contemporary society understands and interprets its past. Because that past is about the lives (and deaths) of human beings, this responsibility is a moral one. It cannot be served in a medium which removes, or immolates, historians from the public sphere
- there is no accreditation structure or overall ‘governing body’ of historians. The edifice of historical learned societies in the UK and abroad is national and local, sometimes specific subject-based. Although the Royal Historical Society, Historical Association, and Society of Antiquaries often speak on behalf of the profession to government, they do not claim to represent this more complex world of learning.

2. Challenges and Opportunities of ICT in Historical Research

I concentrate on the following areas where ICT has proved, or is proving, significant in historical research over the past five years or so. I have divided them (somewhat arbitrarily) into six areas:

1. Information search, retrieval and validation
2. Availability of historical raw materials in facsimile forms
3. Publication and accessibility of textual and non-textual historical outputs
4. Processing/collation of large volumes of data
5. Maintaining formal, institutional and semi-institutional contacts in the world of learning
6. Fostering informal research contacts around particular research themes and objectives

1. *Information search, retrieval and validation*

This is perhaps the most striking and immediate impact of ICT upon research practitioners in history. As in the wider non-academic community, the current preferred instrument of general internet search is Google. Historians are not particularly worried about the inevitable redundancy in the search results; nor about the non-hierarchical and flat nature of the results. They are relatively comfortable with their capacity to evaluate the reliability of a particular hit. The skills required replicate those that historians use in the evaluation of their historical evidence as a matter of course. But library and archive catalogue searches may well rival Google searches for many practising historians. Union catalogues, or their equivalent in the archive domain (COPAC, A2A, national library catalogues abroad) are essential resources in some historical fields for information search. Historical Abstracts, American History and Life, and the Annual Bibliography of British and Irish History are essential ancillary bibliographic tools to studying history in the relevant subject areas, or within English-language publications. The lack of equivalents in history for other language areas is increasingly crippling to their success in the modern historical research economy. Important, too, are some specialist information providers (e.g. Wellcome Institute for medical history). There is no one information gateway for historians. We are used to a world in which we expect to develop quite sophisticated information literacy about where we are most likely to find relevant information. We are eclectic and philandering, inherently suspicious of monogamous relations with information providers (and tend to equate monogamy with monopoly, or the possibility of withholding information). Information search and retrieval often provides, however, an 'institutional' or even 'received' wisdom about a subject. Historical research, at its best, is subversion of 'received' wisdoms about the past. In that sense, information search in public domains about historical matters is either about confirming details, or (if about larger questions) generally only the beginning of a historical enquiry, and not the end of it.

2. *Availability of historical raw materials in facsimile forms*

The availability of historical materials in facsimile has been often driven by commercial forces, the vagaries of research grant submission success rates, and the activities of pioneers in the field. The result has, by and large, not been 'strategic' (but could it have been?). But it has not been bad. The big success story is STConline and its 18thC successor. It is not an exaggeration to say that these instruments have transformed the nature of the questions that historians ask (and can expect an answer) of their materials. They have had a demonstrable 'cyclonic' effect on their research areas. ICT has been called an essentially 'disruptive' technology. Its effect here, in terms of the ability to get answers to questions very quickly, speeds up the iterative process at the heart of the historical process. It also democratizes the historical processes as well. We might expect the same effects wherever facsimile historical raw materials have been created. In reality, however, the impact has been more variable – registering the great differences in scale, accessibility, searchability, and relevance of historical raw materials to solving one or more kinds of question. We should note that:-

- some worthwhile historical raw materials in ICT facsimile forms are, in reality, only accessible currently with difficulty
- there is a mismatch of provision. The areas where the largest amount of historical raw material is available in traditional forms is in the post-1800 historical domain. This is also where the majority of research practitioners' interests in UK HEI are also located. This is the area, on the other hand, where historical raw materials in facsimile forms are least commonly available. Those that are available (e.g. the Census data) have been created for a number of reasons (public policy; social science research; genealogists' needs, etc), and historians' research interests have not often figured large among them
- in common with other disciplines, historians are faced with a situation where interoperability and cross-searching of one historical dataset with another is becoming a major stumbling block to their being effectively deployed to undertake historical research
- historians, in common with other arts and humanities disciplines, have benefited from the availability of non-textual historical raw materials in facsimile forms. But this raises particular issues of fitness for purpose, and whether a facsimile digital representation really serves as a surrogate for historical research in e.g. art history or material culture
- historians are still undecided about the cost-benefits of full facsimile reproduction over more selective cataloguing and calendaring. The equation is not one of universal applicability, and past experience is no guarantee to likely future need and *praxis*. It is likely that, for the foreseeable future, no matter what the subject domain or the volume and sophistication of the historical materials available online, historical practitioners will still rely on independent field-work in archives, libraries and repositories for their most significant research conclusions

3. *Publication and accessibility of textual and non-textual historical outputs*

In common with other disciplines, historians have begun to take advantage of publishing historical outputs in electronic media. The process has generally been one in which the initiative has been with the information mediators – the traditional publishers making e-books and e-offprints available, journal editors making backruns of journals available through JSTOR, etc. Although the open access movement has provided additional facilities, I am not aware that historians have made extensive use of them. The problems in this area – the need to provide publications that are of high quality, consistently available, methodically catalogued, peer-reviewed, where the access to the material is guaranteed by reliable protocols - are not unique to historical studies. They are none the less relevant to us than to other areas of scientific study. Because of the historians' public role, however, they might be even more relevant, and we should expect electronic publication to be an arena in which there will be more high-profile public contestation over historical issues in the short to medium-term future.

4. *Processing/collation of large volumes of data*

We have already (implicitly) outlined three different layers of historical ‘stuff’ available in electronic form:-

- information of various kinds, catalogued and uncatalogued, professionally and unprofessionally produced, mediated and unmediated
- historical ‘raw’ material in facsimile form of various kinds
- works of history available in electronic media

There are probably other layers too.

The reality is that the application of historical intelligence (the creative process at the heart of any historical field of study) requires access to all three concurrently, and in a fashion that cannot be predicted or pre-determined. So the storage of digital information in many different formats, stored and managed in bespoke systems by independent institutions, often does not make the application of historical intelligence through ICT significantly easier than in traditional media. Some significant historical data is only available in particular institutions through commercial site-licences. Other data has to be purchased on an item-by-item basis. Even when available, its manipulation or accessibility can be restricted by the terms of the particular license in question.

Historical material is often rebarbative. It is often not readily capable of being analysed in statistical fashion. The loss of granularity in converting a historical document into a numeric or field-delineated dataset is often a limitation to its historical analysis, and certainly to the subsequent re-usability of the dataset in question. Where statistical methodologies have been adopted, they have often resulted in very sophisticated and important conclusions (e.g. historical demography: social history; crime, etc). But they have been accepted because those involved have the specialist skills both to understand the documentation involved, and to apply the social-science techniques to the analysis of the material.

Historians have been reluctant to develop over-arching ontologies to their subject domain of a kind that would facilitate cross collection searching. That is because historical conceptualization is fluid, often essentially contested. At the level of conceptualization, it would be generally regarded as counter-productive to attempt to build high-level ontologies for searching purposes. Lower-level ontology development is still in its infancy.

Historians are consummate list-makers. Lists constitute the building blocks for our contingent spatial, temporal and nominal historical awareness. But such lists tend to remain private, implicit and subsumed into the historical research process. The historian of eighteenth-century London might, for example, have a working list of goldsmiths at work in the capital in the eighteenth century, and use that in the course of an analysis of its social history, or the operation of credit. But that list might not be explicit in the eventual published work, and it would certainly not be electronically available, even though it might be of considerable research utility to others. There is an issue of what constitutes a ‘publishable conclusion’ and a ‘historical dataset’ here which needs to be addressed.

5. *Maintaining formal, institutional and semi-institutional contacts in the world of learning*

Historians have benefited from the large number of institutional structures that exist to support the historical world of learning. AHDS History is strong, responsive to needs, and innovative. The TNA has also been leading the world of archivists, a particularly well-organised group in the application of ICT to records management. The BL is a world-leader in both library and archive ICT applications. The UK has a strong infrastructural provision for historical studies, in comparison with many European neighbours (for example). But there is an issue about coordinating these various institutions, and linking them to the work of historical studies, especially when their mission statements (in some instances) require them to be responsive to public demands in other ways. Although we are not directly involved in the funding of these bodies, we should be aware that their existence is dependant on the use that we make of them. That use implies being able to demonstrate benefit, directly or indirectly from their role.

6. *Fostering informal research contacts around particular research themes and objectives*

The world of historical studies is composed of largely self-reflexive groups of scholars and practitioners, whose activities and interests touch and intersect one another at a variety of points, which constantly change and evolve with the research trajectories and careers of particular individuals. They reflect fashions and trends in historical enquiry, and respond to dominant personalities and the formation of research clusters in particular places. In this respect, historians are no different from other constellations in the republic of science and letters. There are some examples of historical practitioners making use of VRE [Virtual Research Environments]. But I suspect that the concept, and certainly the results of those experiments, is not generally in use as yet in the historical community. This may partly be a result of an important practical constraint on the application of ICT in the historical domain. Most academic departments in history do not have regular access to skilled technician time. History is a popular undergraduate subject with correspondingly high staff-student ratios. There is limited time for individuals to equip themselves with the necessary technical skills – and relevant courses are generally not available for them to do so. What technical proficiency they have acquired so far has been ‘on the fly’. They have become used to that being the norm. Technical developments that are going to assist their research *praxis* have to be weighed against that background.

3. **Historians and the E-Science Agenda**

E-Science – in this context it might be better to call it E-Research – is flavour of the month. It means all things to all people. I am sticking to the definition of the National E-Science Centre:-

in the future, e-Science will refer to the large scale service that will increasingly be carried out through distributed global collaborations enabled by the Internet.

Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists.

[<http://www.nesc.ac.uk/nesc/define.html>]

How, then, should historians react to this possibility?

To answer that question we need to focus on these issues. These emerge from this analysis of how historians work, and how they have reacted so far to ICT opportunities and challenges:-

- We have some fairly large datasets, and some potentially very large demands for fast transfers of data (high resolution of images; video and sound footage of historical materials; national census data, etc). How should we identify where these are, and where their further study and exploration in an e-science Grid environment can be most profitably enhanced?
- We have some potentially very disparate data and some emerging interoperability needs. How can the Grid help us with these? Can we square the challenges of the authentication, licence, copyright and accessibility issues with the opportunities that potentially may lie ahead for us? If so, can we identify a particular domain or area of enquiry where we should concentrate our attentions?
- If we can, what kind of interoperability are we looking to provide? If a high-level ontology of the subject is unlikely to exist in the near future, and perhaps ever, does the provision of lower-level ontologies meet the needs of interoperability? Or are there other ways in which we can develop a Grid-based 'research platform' that would more appropriately provide for historians' needs?
- We have a strong individualist research culture; but also proven evidence of successful collaborative endeavour. Can we base our Grid developments on the latter? If so, what lessons should we draw on so that we can maximise the likely outcome? Are there changes that we need to contemplate in our research culture? Do we need more (or more advanced) research tools? Are these likely to be generic to the arts and humanities, or specific to history?
- We have important infrastructure partners in libraries, archives, etc. The latter, however, are (for the most part) not leading players in the Grid and many lack the necessary technical infrastructure. Is there a potential asymmetry here that we need to address?
- Our informal means of networking are effective in environments where technical support is minimal. How should we approach the move to an environment of higher technical specification for informal networking, with its greater demands? What is the best way of ensuring that historians make effective use of the Access Grid? Of VREs? How should we approach the issues of awareness, training, and support?
- Research collaboration poses fundamental questions about the 'ownership' of research outputs (in a world in which the RAE morcellises those into a return by each individual practitioner). Historians are both very willing to share research results, and quite jealous of their particular 'subject' and sometimes of the 'archive' to which it relates. There is an issue here about the cost-benefit of

specialist knowledge of a particular subject in the hands of one individual, and the collaborative understanding of a historical domain (which, because of the nature of the subject, may never be a completely 'shared' understanding).

Historians, like every other scientific discipline, are faced by an explosion of information. E-science is proposed as a means of managing that data 'bonanza'. The prospect, the overwhelming rationale for this seminar, is that e-Science methodologies help to raise our levels of abstraction (in historical terms, our levels of 'historical intelligence'). If this is so, we should be looking seriously to take advantage of what the e-Science agenda has to offer us.

M. Greengrass
6 June 2006

Appendix Three

Institutional Profile of the Historical Discipline in 2001

The following table is based on an unpublished analysis of the returns in the 2001 RAE exercise:-

AHRC Subject Panel 2001 RAE Subjects Below	Totals No. Individuals Submitting to RAE	Institutional Size By Staff Numbers/ Total Number of Institutions With Departments			
Medieval and Modern History 1 History – 1077.9 2 American Studies – 113.5	1833.4	1	50+ 30-50 10-30 -10	2 10 49 32	93
		2	50+ 30-50 10-30 -10	-- -- 4 9	13

Source: HESA
