

Tools and Methods for Historical Research

A Methods Network Working Paper

The designation of this document as a 'working paper' is an acknowledgement that its content is not meant to be regarded as finalised or fixed. As part of the Methods Network remit to encourage discussion of the advanced use of ICT tools and methods for arts and humanities research, an open forum is being setup where this paper – and all matters relating to digital historical research – can be freely discussed. Please see the Methods Network website for details.

Introduction

The objective of this working paper is to provide some examples of the type and range of digital tools (and associated methods) that might profitably be used by researchers working in the discipline of History. In keeping with the remit of the Methods Network - one of the objectives of which is to promote the dissemination of knowledge about ICT related tools and methods across subject boundaries - some reference will also be made to techniques and tools normally associated with other disciplines which clearly have applications to historical research.

On a theoretical map of arts and humanities subject areas where disciplines could be positioned to try and demonstrate overlap and methodological common ground, there would be a strong case for placing the discipline of History squarely in the centre of the map, providing as it does a chronological backbone for all areas of research, as diverse as musicology, practice-based art, theology, literary studies and so forth. In terms of immediate adjacency, the borders of what might be considered 'historical research' would blur imperceptibly with a number of neighbouring subject areas: archaeology, classical studies, sociology and art history for instance; all of which have tools and techniques associated with them which are (or potentially could be) of use to historians. Appropriated techniques from computing and information science such as data mining and e-Science related methods are also increasingly coming to the attention of those seeking to explore new methods of engaging with research questions - as are techniques associated with geographical mapping and information systems - so it is clear that the scope for employing a variety of technical methods is very broad.

This view is echoed in a recent report¹ written by Onno Boonstra and colleagues who outline numerous methodological and thematic areas of potential research, many of which they feel have yet to be tackled in a coordinated and purposeful way. Whilst very useful as a roadmap for possible engagement with a wide variety of technological tools and methods, the wider context of the report is gloomy in its overview of the recent state of 'historical information science' (the authors' preferred term for computer-related historical research). At the time of writing in 2004, they expressed the view that the results of nearly two decades of History and Computing were 'slightly disappointing' and went on to qualify this statement as follows:

They are not disappointing because 'computing' failed to do what it intended to do, which was to provide 'history' with computerised tools and methods historians could use to expand the possibilities and to improve the quality of their research, but because 'history' failed to acknowledge many of the tools 'computing' had come up with.²

One of the major challenges that the authors identify is the provision of an effective internationally accepted forum for discussing and disseminating the results of this kind of research and it is clear that the various

¹ Boonstra et al (2004)

² Boonstra et al (2004)

iterations of the Association of History and Computing (ACH) have a critical role to play in coordinating and promoting these kinds of activities,³ both at international and national levels.⁴

Tools and Web Resources

To address the question of what is currently available to historians, one must inevitably turn first to the Web for an overview of initiatives that are the result of recent (and not-so-recent) attempts to use digital tools to break new ground in areas of historical research. One early and very influential project that is widely cited as providing a benchmark for presenting complex historical research material on the web is the 'Valley of the Shadow',⁵ a hypermedia archive about northern and southern communities during the American Civil War containing more than 100,000 items taken from newspapers, letters, diaries, photographs, maps, church records, population census data, agricultural census data and military records. Hosted by the Virginia Center for Digital History, this project might usefully represent the medium of the web itself as one of the most ubiquitous and most widely accepted tools, complementing the use of email as the other major communication tool that presumably all historians now accept as an intrinsic part of their activities.

A number of repository/gateway/portal/listings sites exist to aid historians in finding relevant resources (see below for a representative sample of these sites – all sites active 1 March 2007).

- AHDS History, large searchable data archive which also contains documents about deposition of material and best practice information, <http://ahds.ac.uk/history/>
- UK Data Archive, the largest collection of digital data relating to the Social Sciences and Humanities in the UK, <http://www.data-archive.ac.uk/>
- University of York, Guide to Using Historical Resources on the Internet, <http://www.york.ac.uk/teaching/history/pjpg/internet.htm>
- Intute, Internet for Historians, <http://www.vts.intute.ac.uk/he/tutorial/history/>
- Digital Librarian, A Librarian's Choice of the Best of the Web, <http://www.digital-librarian.com/history.html>
- University of London Research Library Services, History, <http://www.ul.ac.uk/subjects/guides/historytools.shtml>
- Institute of Historical Research, Best of the Web, Internet Resources for History (Humbul) http://www.history.ac.uk/guides/internet_historians.pdf

With such a plethora of very effective portal-type tools for finding what amounts to an enormous array of resources for historical research, it is useful to have a recent AHRC funded report entitled 'Peer review and evaluation of digital resources for the arts and humanities'⁶ which is based on a total of 777 responses to a survey circulated to a number of (mostly history-related) mailing lists. 442 responses were received to question 6 which asked respondents to:

Name the three digital resources which you use most often in your own research (such as TLG, EEBO, RHS Bibliography, Old Bailey Proceedings, Oxford DNB; please do not include online journals, library catalogues or search engines). Why are these resources of such value?

³ Association for History and Computing, <http://grid.let.rug.nl/ahc/intern/index.html>, (active 9 March 2007); see also, Association for History and Computing (2005 conference), <http://www.ahc2005.org/en/>, (active 9 March 2007)

⁴ See also: UK branch of the AHC, <http://www.gla.ac.uk/centres/hca/ahc/>, (active 9 March 2007); and AHC in the U.S., <http://www.theaahc.org/main.htm>, (active 9 March 2007)

⁵ University of Virginia, The Valley of the Shadow, <http://valley.vcdh.virginia.edu/> (active 1 March 2007)

⁶ Institute of Historical Research, Peer Review and Evaluation of Resources for the Arts and Humanities, http://www.history.ac.uk/digit/peer/Survey_report2006.pdf (active 1 March 2007)

The top ten resources listed were as follows:

Resource	Responses
Oxford Dictionary of National Biography	122
Royal Historical Society Bibliography	77
Early English Books Online	58
Eighteenth Century Collections Online	28
Times Digital Archive	24
Archaeology Data Service	19
The Proceedings of the Old Bailey	18
British and Irish Archaeological Bibliography	12
British History Online	11
Canmore	10
Historical Abstracts	10
Thesaurus Linguae Graecae	10

Table 1 Taken from Peer Review and Evaluation of Digital Resources for the Arts and Humanities (pg.5)

Whilst these ‘resources’ are slightly to one side of the main thrust of this paper - the principal purpose of which is to consider ‘tools’ - it is nonetheless illuminating to understand which sites are perceived as particularly valuable. In addition to *relevant content* which will almost always be the principle reason that users will keep on returning to specific websites, a less important but nonetheless significant enticement will be:

- The functionality that is available (e.g. search and sort options)
- The flexibility of the data (e.g. variant spelling detection; text hyperlink to original manuscript image)
- The robustness and durability of the data (e.g. the technological platform; the encoding standards employed)

Taking some of the above resources as examples, it may be useful to consider some of the tools associated with their development, functionality or implementation to understand what elements other historians might usefully aim to employ in their own ICT-related resource development projects.

The Oxford Dictionary of National Biography uses a very flexible search interface which offers a ‘type and hope’ simple search field for finding a specific person or word from the full text of the DNB, but also offers five other search screens incorporating many more fields (sometimes linking to further ‘advanced’ options e.g. name search) which allow the user to access all manner of gender related, professional, geographical, chronological, religious and financial information – in addition to image and bibliographical data. One of the comments in the survey pointed out the specific usefulness of this site in creating prosopographical statistics and it is clear that the online DNB fulfils the criteria often quoted for establishing the legitimacy of a technological resource, i.e. that it allows researchers to access information that would be impossible or very difficult to have discovered before that resource became available.

Obviously, one of the critical tasks in assembling a resource of this magnitude (55,800 lives; 63 million words; 10,300 portraits) is the rendering of the original text into a machine-readable format, and whilst this is less of a problem for text that has incrementally been transferred into digital formats over the years (which may have been the case with the DNB) it becomes a huge technological challenge when major repositories of primary source material are only in printed formats and require either manual text transcription or some form of scanning procedure (image and/or optical character recognition) to transform them into digitally accessible archives.

The Eighteenth Century Parliamentary Papers Project,⁷ one of the initiatives funded by the Joint Information Systems Committee (JISC) digitization programme, will use very fast bulk scanning equipment that is capable of scanning 600 pages per hour, facilitated by vacuum enabled page turning technology and laser guided edge detection sensors. The size, weight and cost of this equipment (see fig.1) clearly requires a long term institutional commitment and will exceed the strategic requirements of many organisations and projects, but it is symptomatic of a new approach to resource creation that such tools now exist that can process around one million scanned pages a year, using techniques that are sensitive to the fragile nature of the primary sources and mindful of the technical standards intrinsic to sustainable resource creation. At a recent Methods Network workshop,⁸ one participant reported that the Open Content Alliance is using this kind of hardware in a multiple-machine facility that perhaps begins to approach the sort of capabilities available to commercial entities such as Google and Microsoft, who are separately in the process of creating colossal digital repositories of books numbering millions of items.



Fig. 1 University of Southampton, BOPCRIS⁹

Whilst different in scale to the commercial initiatives referred to above, Early English Books Online (EEBO)¹⁰ is an important archival record of virtually all books produced in Great Britain, Ireland and British North America before 1700, and also features prominently in the AHRC Peer Review Project (see Table 1). The relevance to historians of this archive extends to an enormous range of subject matter that encompasses royal, governmental and provincial official public documentation as well as a prodigious amount of social historical material relating to all classes of society. Whilst the image scanning of book pages in this project is of enormous significance, it is the related Text Creation Partnership (TCP) initiative that is of further interest in the present context. EEBO-TCP¹¹ is managed by the Universities of Michigan and Oxford and financially supported by over seventy other institutions to produce manually keyed and SGML/XML encoded text editions of a significant portion of the EEBO corpus. The aim is to create full text editions of about 20% of the works represented by images in the archive, thereby enabling search

⁷ British Official Publications Collaborative Reader Information Service (BOPCRIS), 18th Century British Parliamentary Papers, <http://www.bopcris.ac.uk/18c/> (active 1 March 2007)

⁸ Methods Network, Open Source Critical Editions Workshop, <http://www.methodsnetwork.ac.uk/activities/act9.html> (active 1 March 2007)

⁹ Joint Information Systems Committee (JISC), News, http://www.jisc.ac.uk/news/stories/2004/07/project_second_in_world_news.aspx (active 1 March 2007)

¹⁰ Chadwyck Healy, Early English Books Online, <http://eebo.chadwyck.com/home> (active 1 March 2007)

¹¹ University of Michigan, Text Creation Partnership, http://www.lib.umich.edu/tcp/eebo/proj_des/pd_intro.html (active 1 March 2007)

functionality that goes way beyond keyword-type linguistic analysis processes (e.g. concordance, collocation and clustering), highly useful as those are for historians as well as those involved with diachronic linguistics. As a result of the textual encoding though, searches can also be formulated that, for instance, only search for proper names, or only search in the marginalia, or perhaps only search for non-English terms that appear in stage directions.

The decision by the Text Creation Partnership to rely on laborious methods of manual keying and encoding clearly has a critical effect on what it is possible to achieve given certain funding criteria, and is in contrast to the abovementioned Eighteenth Century Parliamentary Papers project which declares on its website that Abbyy OCR (optical character recognition) software will be used in conjunction with the Agora Content Management System to automate as much of the data management as possible. Given that there is no single recommended solution for major projects like these, it is a useful reminder that the best strategic methodologies and the most appropriate tools need to be defined according to the nature of the work being undertaken. This might encompass cases where a hybrid solution involving automation combined with human expert feedback into the system is the preferred methodology, improving the chances of more accurate scanning as the project progresses and as the system learns from its mistakes. The Gamera project¹² is an open source initiative to supply a programming framework for building recognition systems for 'difficult' historical documents. It is envisaged as a tool that programmers will use in conjunction with subject experts and supports a reiterative test and refine model, a graphical user interface and a modular plug-in architecture that is capable of performing five separate document recognition tasks:

- pre-processing
- document segmentation and analysis
- symbol segmentation and classification
- syntactical or structural analysis
- output

Gamera stores output in an XML-based file format, which enhances its interoperability with other systems and is also in keeping with its open source ethos.

Database Structures

Of all of the computer-related tools for historical research, database (and statistics) packages have had the most consistent use by the widest constituency of researchers using digital methods in the discipline. In general terms, approaches in the past generally focused on information originating from social science research and the data that resulted from this work often took the form of lists. As such, the manipulation and analysis of this data (e.g. census records, state finance data, demographic data, mortality data etc.) leaned itself primarily to being handled by database packages and there is a considerable amount of literature devoted to discussion of the different approaches to structuring data and the use of relational database management systems (RDBMS)¹³, such as dBase¹⁴, Paradox, Clipper, FoxPro and Microsoft Access. One of the principle areas of discussion within this area was how to most usefully represent historical data and much of this debate seems to have focused on the relative merits of a 'source oriented' versus a 'model oriented' approach to structuring the information.

The 'model-oriented' structure required a great deal of preparatory data analysis and provided logical organizational entities into which defined data elements could be deposited (i.e. database fields). At the outset of a project then, not only was the researcher obliged to work out a sensible way of dividing the

¹² Johns Hopkins University, A Gentle Introduction to Gamera, <http://dkc.jhu.edu/gamera/html/overview.html> (active 1 March 2007)

¹³ For a detailed bibliography on database literature (pre-1996) see Harvey & Press (1996)

¹⁴ dBase is often referred to as a 'relational' database even though it doesn't meet the criteria defined by Edgar F. Codd's Relational Model

information into rational and discreet entities, but he or she also needed to think very carefully about what sort of information would need to be extracted from the system during its working lifecycle. This model was particularly effective in cases where the data was relatively simple, complete and regular in its structure.

The 'source-oriented' approach attempted to more comprehensively represent the source information *in its original form* using a combination of textual markup and fielded data structures, thereby capturing as much of the subtlety and nuance of the original data as possible and also enabling future researchers to formulate their own retrieval and analysis approaches to the information. As a leading exponent of the 'source-oriented' camp, Manfred Thaller developed the KLEIO¹⁵ system which represented the original source data at two levels: firstly as uninterrupted strings of arbitrary characters and secondly, as meaningful units such as numerical or calendar data etc. The more formally structured element of the system, arranged hierarchically and referred to as the *knowledge or logical environment*,¹⁶ accurately referenced the transcription (i.e. the full-text) layer, and all queries made of the transcription layer were interpreted via that logical environment – the net advantage being that the researcher still had a 'machine-readable' version of the original to refer to.

Whilst contextually useful for talking about data modelling, the debate about 'source' versus 'model' oriented approaches – which very broadly speaking might translate into 'encoding' versus 'database' - has diminished in importance due to the potential of XML query processes (e.g. Xpath and XQuery) and the relatively recent acceptance of the term 'XML database' to denote data in XML format that is interoperable with relational database packages such as DB2, Oracle and Microsoft SQL Server. Where the functionality of a database is required, projects will undoubtedly continue to incorporate them into their technical strategy, but after a long period of standards being discussed intensively, it is clear that acknowledgement of some form of XML data interchange mechanism is almost a *sine qua non* of project funding, mainly because of concerns about preservation issues but also because of the growing awareness of the need to have an effective way of globally harmonizing disparate information systems.

Data Mining

Having considered issues to do with the presentation and structuring of data, the former principally in the form of web resources and the latter as database and textual information systems, the next stage of the information cycle involves the querying and analysis of that data. One deceptively simple experimental project, H-Bot,¹⁷ is an example of a 'question answering' (QA) system which accepts natural language questions and attempts (in one of its modes) to return exact answers using the Google index to find statistically likely matches on the key keywords that the user has entered. The system is capable of performing reasonably well in the context of a certain type of factual question and its creators argue that further development work on the underlying rule sets would further improve its capabilities. Designing QA systems, according to Dan Cohen, 'exercises almost all of the computational muscles',¹⁸ involving as it does: search methodologies, document classification, question interpretation (natural language processing) and statistical and linguistic text analysis.

Cohen has also used a similar approach with another data mining system called Syllabus Finder¹⁹ which uses a keywords-in-context (KWIC) approach to return highly relevant results relating to syllabi information. Having determined a relevant keyword set using word frequency analysis from a known set of documents related to syllabi, the search term entered by the user is optimised by the addition of these terms and the

¹⁵ University of Göttingen, The Principles of Kleio, <http://wwwuser.gwdg.de/~mthalle2/manual/tutorial/intro.htm#mark0> (active 1 March 2007)

¹⁶ For an explanation of the KLEIO hierarchy, see Denley (1994)

¹⁷ George Mason University, Centre for History and New Media Tools, <http://chnm.gmu.edu/tools/h-bot/> (active 1 March 2007)

¹⁸ Daniel J. Cohen, D-Lib March 2006, <http://www.dlib.org/dlib/march06/cohen/03cohen.html> (active 1 March 2007)

¹⁹ Daniel J. Cohen, D-Lib March 2006, <http://www.dlib.org/dlib/march06/cohen/03cohen.html> (active 1 March 2007)

bundled query then interrogates Google's API service (as well as a locally stored database) which returns search results in a SOAP envelope (an XML schema used for server-to-server communications). Additional statistical and expression matching analysis is then carried out on this dataset resulting in highly relevant and specific output that is capable of identifying college or university organisation names and course assigned book titles. These and other tools created at CHNM are complex in their conception but are designed with simple interfaces and have the feel of small group developmental research systems that have been constructed to address particular user-defined needs.

A more widely collaborative data mining initiative that is currently focused on information retrieval in the area of literary studies is the NORA project,²⁰ the objective of which is 'to produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries.'²¹ The development of this software is being done by a consortium of organisations involving subject specialists and computing scientists and the functionality of the demo version of the system displays a similarly distributed methodology. The user interface software runs as a Java Webstart application delivered to the user via one server; the relational database management system (Tamarind) runs on another server; and the D2K data-mining framework²² currently runs on a third server. The D2K (Data-to-Knowledge) framework supports other projects: T2K (Text-to-Knowledge) and M2K (Music-to-Knowledge), all of which are being developed to play a significant new role in how the Humanities Computing community works together to build, use and share tools.

The principle is that 'modules' of well defined reusable code can be implemented as single or nested components to carry out a wide variety of functions – which in the case of the T2K framework would include a rich set of natural language pre-processing tools which could carry out: lemmatization, tokenization, part of speech tagging, data cleaning and named-entity extraction.

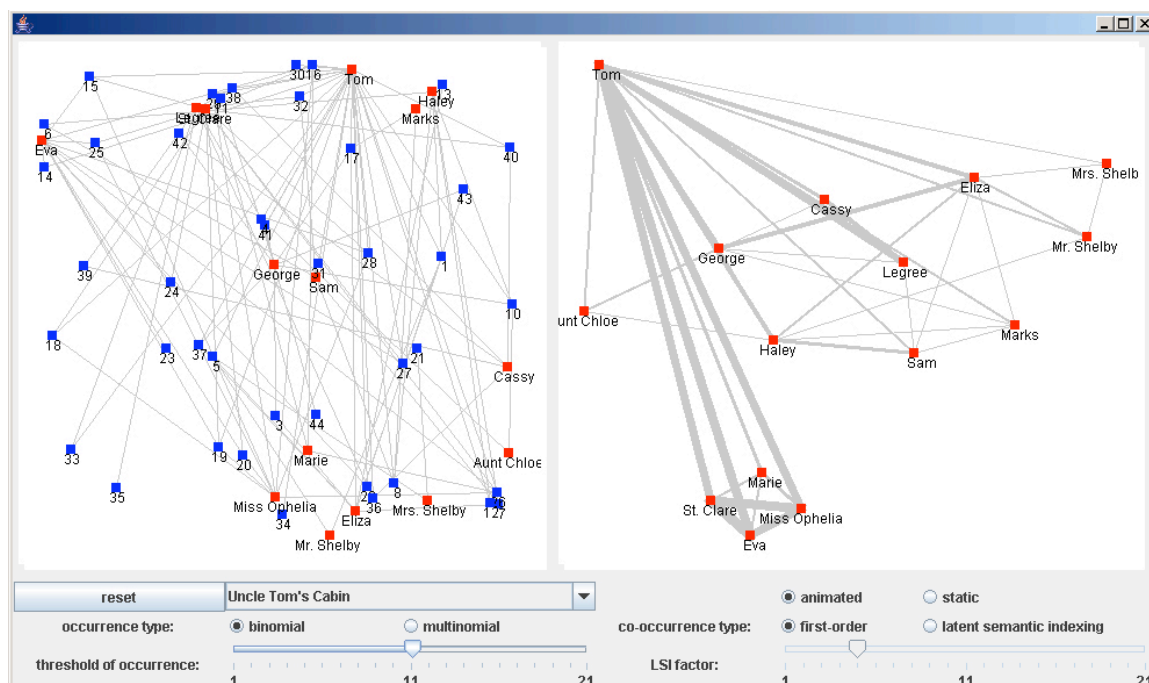


Fig 2 Screenshot from the 'Social Network Demo' available from the Nora Project

²⁰ The Nora Project, <http://www.noraproject.org/> (active 1 March 2007)

²¹ The Nora Project, <http://www.noraproject.org/description.php> (1 March 2007)

²² University of Victoria, D2K Mining, http://mustard.tapor.uvic.ca/cocoon/ach_abstracts/xq/xhtml.xq?id=201 (active 9 December 2006)

In the context of the NORA project, the concept is to build an application using the D2K framework which will retrieve 'only what's needed' from large digital repositories (of XML encoded text) and place that information into a database, thereby cutting down on the processing overhead of having to deal with large amounts of redundant or null data. Once collected, an additional function of NORA is then to provide visualizations of the analysis of that data that will provide researchers with clearer insights into a range of issues including social networks, content overview, document classification types etc. (see Fig. 2).

A project based at Sheffield University is also looking at the problem of mining information intelligently from distributed data sources and is specifically addressing how to deal with the sort of fuzzy and ambiguous data that often characterises historical information. The ARMADILLO project²³ is based on twelve datasets that contain information about eighteenth century London and the objective is to employ probabilistic disambiguation methods (using pre-determined algorithms) to the heterogeneous data, which is then mapped to a fairly simple ontology consisting of the following entities:

- Name
- Role
- Place (i.e. place name which may be subject to a change of location)
- Location (associated with a map reference)
- Time (either point in time or time period)
- Resource

Users would then be able to turn these categories 'on' or 'off' depending on the nature of their search and the likelihood of returning meaningful results, allowing for much finer control of the data retrieval process. The stated aim of this project is to employ as much 'knowledge' to the task of initially processing information as possible, in order that the likelihood of error introduced by synonyms, homonyms, variant historical spellings, inconsistencies introduced by human error - and so forth - are minimised when records are retrieved for analysis.²⁴

Quantitative Methods

The use of quantitative methods is an intrinsic part of historical research but has not always met with blanket favour over the years in all areas of the discipline. In a recent essay, William Thomas²⁵ cites the controversy over the publication in 1974 of Robert Fogel and Stanley Engerman's *Time on the Cross: The Economics of American Negro Slavery* which (in amongst a vast amount of other detail) posited the claims that only two percent of the value of income produced by slaves was expropriated by their masters and that the typical enslaved person received less than one whipping per year. These arguments were based on economic models and mathematical methods of analysis (cliometrics) and unsurprisingly met with fierce criticism from many quarters. Despite this sort of opposition, however, the use of statistical techniques is deeply embedded into economic and social history research, as well as being a central component of data mining. In journals such as: *Social Science History*; *Social History*; and *Economic History Review*, accounts of the methodological processes employed in the various analyses featured are subordinate to the factual historical substance of the article and may therefore be giving a misleading impression of the acceptance and use of these techniques in various sections of the historical research community.

The most widely used statistical tools are borrowed from the Social Sciences where there seems to be a consensus that the SPSS²⁶ system, first developed in 1968, continues to offer the correct balance of

²³ Humanities Research Institute, Armadillo Historical Data Mining, <http://www.hrionline.ac.uk/armadillo/sources.html> (active 1 march 2007)

²⁴ A podcast of a presentation on this project is available at: Methods Network, <http://www.methodsnetwork.ac.uk/resources/podcasts.html> (active 1 March 2007)

²⁵ Thomas (2004)

²⁶ SPSS, <http://www.spss.com/> (active 1 March 2007)

usability and functionality for most forms of analysis. Minitab²⁷ software is also widely referenced by course descriptions for academic departments offering applied statistics modules, as is SAS,²⁸ despite the fact that there are also numerous freeware and open source options available, many of which claim to focus on specific areas of functionality.²⁹ The ways in which statistics packages can be used is a complex and highly specialised field of study in its own right, but Boonstra et al propose the following methods as holding 'great promise for future historical research.'³⁰

- *Logistic Regression* – a model for predicting a dual category variable (e.g. married/unmarried, live/die, north/south)
- *Multilevel Regression* – an analysis that includes hierarchical data (e.g. data about patients, the doctors treating them, the hospitals they work in, and the region the hospital is located)
- *Event History Analysis* – a study of the independent variables which may contribute to the likelihood of an event occurring
- *Ecological Inference* – reconstructing individual behaviour from aggregate data where there is a paucity of information relating to the individual
- *Time Series Analysis* – a range of techniques for studying change over periods of time

Boonstra et al contend that the adoption of these techniques across all areas of historical study would enable interesting and productive research.

A related area where quantitative tools are widely used is in the construction of simulations, another technique that has been in existence for a very long time but has not been taken up consistently across the community. The purpose of this technique is to analyse behaviour and events in the context of historically given variables in order to gain a better understanding of cause and effect within a defined system. Very old models include the semi-computerised attempt to simulate the outbreak of World War One by Hermann and Hermann³¹ (1967), and also SOCSIM³² (1970 onwards), a demographic micro-simulation program that enabled probabilistic analysis of marriage and fertility patterns within a closed society. Despite the lengthy history of their use in historical research, the application of current simulation techniques largely appear to be associated with risk assessment for businesses, environmental planning and scenario reconstructions for military situations.³³

Markov Chain sequences, which are widely used in simulation models and are relevant to determining the probability of an event occurring based on the occurrence of a previous event, have also been very influential in the field of Linguistics where much research has focused on probability models for automatically tagging large corpora. It should be apparent that the potential for corpus linguistics techniques to impact on historical research is significant and that problems such as the disambiguation of pre-modern-era spelling variants (which is the focus of the VARD project³⁴ at the University of Lancaster and a component part of the Historical Thesaurus of English)³⁵ should not require duplicate research effort outside of the Linguistics community.

²⁷ Minitab, <http://www.minitab.com/> (active 1 March 2007)

²⁸ SAS, <http://www.sas.com/> (active 1 March 2007)

²⁹ Free Statistics, Free Statistics Software, <http://freestats.altervista.org/stat.php> (active 1 March 2007)

³⁰ Pg. 58, Boonstra et al (2004)

³¹ JSTOR, An Attempt to Simulate the Outbreak of World War One, [http://links.jstor.org/sici?sici=0003-0554\(196706\)61%3A2%3C400%3AAATSTO%3E2.0.CO%3B2-E](http://links.jstor.org/sici?sici=0003-0554(196706)61%3A2%3C400%3AAATSTO%3E2.0.CO%3B2-E) (active 1 March 2007)

³² University of Berkeley, Brief history of SOCSIM, <http://www.demog.berkeley.edu/~wachter/socstory.html> (active 1 March 2007)

³³ King's College London, The Department of War Studies Conflict Simulation, <http://www.kcl.ac.uk/depsta/wsg/consim.html> (active 1 March 2007)

³⁴ Lancaster University, Workshop on Historical text Mining, <http://ucrel.lancs.ac.uk/events/htm06/> (active 1 March 2007)

³⁵ University of Glasgow, Historical Thesaurus of English, <http://www.arts.gla.ac.uk/SESLI/EngLang/thesaur/homepage.htm> (active 1 March 2007)

Visualization

Charles Joseph Minard's visualization of the Napoleonic Army's advance and subsequent retreat from Russia in the 1812-13 campaign is widely cited as one of the most effective graphical depictions of data drawn from statistical sources (see Fig. 3). The thickness of the line indicates the strength of the army during its peregrinations back and forth across Europe whilst the lower graph maps those fluctuations against the recorded temperature on the return leg of the march.

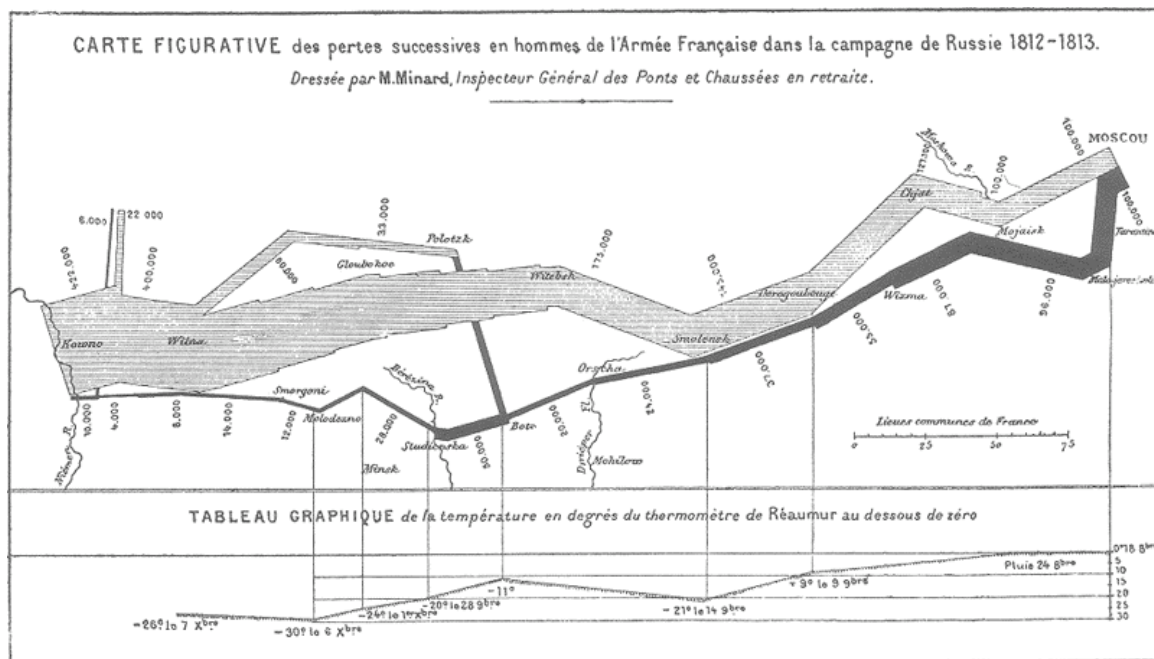


Fig.3 Napoleon's advance across Europe and into Russia

The clarity of information that this graphical display imparts is an impressive model of what visualization can achieve and is a useful benchmark for the type of digital output that it might be possible to create for various forms of historical data.

The Visual Spatial Technology Centre (VISTA) at the University of Birmingham is one of leading centres in the UK for visualization research and one of the many activities that it supports is the Medieval Logistics project.³⁶ During a recent Methods Network meeting,³⁷ Helen Gaffney laid out some preliminary ideas for a visualization of the catastrophic defeat of the army of Emperor Romanus IV at Manzikert in 1071. Using data from: settlement land use; known historical logistics issues and ecological, environment and terrain data, she proposed a visualization that encompassed a variety of aspects of decision theory - including network, game and optimal foraging theories - in order to elucidate the defeat of the Emperor's army at the hands of the less numerous Turkish forces.

Other initiatives in development at Birmingham include a project to investigate and visualize how drapery and adornment in the classical period would have affected movement and interaction between figures; and another demonstration involved the 3D visualization of the contents of a Canopic jar, which is thought to contain a human liver. This kind of work marks out territory shared by historical studies and the cultural

³⁶ University of Birmingham, Medieval Logistics Project, <http://www.medievallogistics.bham.ac.uk/> (active 1 March 2007)

³⁷ Methods Network 'Roadshow' meeting, University of Birmingham VISTA Centre, 1 June 2006

heritage sector and in terms of the cross-disciplinary sharing of tools, it is clear that methods employed by curatorial and conservation professionals, not least in how to describe objects effectively (using tools such as the CIDOC-CRM ontology), could be of enormous value to historians.

The Wroxeter Hinterland Project,³⁸ also based at the University of Birmingham, is a useful example of a project that employed an arsenal of tools to interpret the remains of a Romano-British urban archaeological environment.

An international team of archaeologists is carrying out the total exploration of the city by non-destructive remote sensing methods, including magnetometry, resistivity, electrical imaging, seismic scanning, ground probing radar, airborne hyper- and multi-spectral scanning, and satellite imaging.³⁹

Archaeology as a discipline is generally credited as having had more exposure over a longer period to technological means of research than most other arts and humanities subject areas and as such, represents an enormously useful pool of expertise for colleagues in neighbouring areas of historical research who are looking for new ways of analysing data in their area of research.

Another centre of expertise is the King's College Visualisation Lab,⁴⁰ who are involved with a range of 3D visualization projects that attempt to map internal and external spaces using historical data sources to define verisimilitude as far as possible with the original environments.



© King's Visualisation Lab, King's College London

Fig. 4 Reconstruction of Inigo Jones' Barber Surgeons Anatomy Hall, 1636

Using 3D Studio Max and other tools, stunning results are possible that give fresh insights into how historical spaces were perceived and used and a recent initiative called the 'London Charter'⁴¹ is an acknowledgement that the visualization community requires standardised methods for documenting working practices.

³⁸ The University of Birmingham, The Wroxeter Hinterland Project, <http://www.iaa.bham.ac.uk/bufau/research/wh/Base.html> (active 1 March 2007)

³⁹ The University of Birmingham, The Wroxeter Hinterland Project, <http://www.iaa.bham.ac.uk/bufau/research/wh/Base.html> (active 1 March 2007)

⁴⁰ King's Visualization Lab, <http://www.kvl.cch.kcl.ac.uk/> (active 1 March 2007)

⁴¹ The London Charter, <http://www.londoncharter.org/>, (active 1 March 2007)

At a more accessible level for the individual scholar, devices such as timelines, lexis pencils, graphs and cluster dendrograms⁴² are all examples of data visualization tools.

Geographical Information Systems

To find a remarkable implementation of a system that uses GIS as one of the component parts of its technology, one need look no further than Google Earth to illustrate an essential facet of what it is possible to do using such techniques. Satellite imagery of actual terrain can be overlaid with a wide variety of information choices including transportation, roads, shopping and services information and even user generated information about very specific buildings or neighbourhoods. Obviously, the type of technology that Google has the capability to implement has global significance and requires an extraordinary amount of computational and storage capacity, but nonetheless, the layered approach to information is the same principle that arts and humanities academics (principally perhaps archaeologists) have used for many years when applying GIS principles to research.

The basic unit of representation in a GIS system is the layer, which has information associated with it to describe what that layer represents. *Spatial data* refers to the actual point on the earth's surface that the point or boundaries of the layer refer to and this is described using either latitude and longitude readings or other co-ordinate systems such as the British National Grid or the Universal Transverse Mercator (UTM). The boundaries of the layer for this purpose can be defined as either a point, a line, a polygon, or a pixel, and each of these is represented using co-ordinate pairs. *Attribute data* constitutes the information that is associated with the *spatial data* and might consist of any information, but is generally text or statistics or graphical representations of data that is relevant to that location. The industry standard package is the ArcGIS⁴³ suite of tools which incorporates ArcView and ArcInfo, widely used by academic and commercial project teams, but a range of other products are available,⁴⁴ including a freeware system called GRASS (Geographic Resource Analysis Support System).⁴⁵

It is clear then that GIS has enormous value for the annotation of maps, but it is has perhaps been less clear to historians over the years how a temporal component can be introduced to the system that will allow the description of changes over time. At a recent Methods Network Expert Seminar,⁴⁶ Ian Gregory gave a presentation⁴⁷ that referred to problems of 'time' and 'space' and did so in the context of talking about infant mortality trends in England and Wales from the 1850's to the 1900's. A temporal component to his research was provided by the Great Britain Historical GIS (GBHGIS, Gregory et al 2002), hosted at the University of Portsmouth,⁴⁸ which provides census reports; data on births, marriages and deaths; and on unemployment and poor law statistics, for the entire period of Gregory's study. The present incarnation of this system (in development since 2001) has provided scholars with a new resource in this particular area of research, and has also been the data source for a major lottery-funded project of wider immediate benefit to the public, the Vision of Britain website.⁴⁹ By overlaying statistical representations of infant mortality figures onto precise areas, it is possible to graphically display - in much more detail than was possible before - changes throughout the period and how they relate to a number of factors; notably the difference in death rates throughout urban and rural areas.

⁴² For explanations of these techniques, see Boonstra et al (2004) pp.73 - 80

⁴³ ESRI, <http://www.esri.com/> (active 1 March 2007)

⁴⁴ For a round up of systems see: Virtual Terrain Project, <http://www.vterrain.org/GIS/index.html> (active 1 March 2007)

⁴⁵ Geographic Resource Analysis Support System (GRASS), <http://grass.itc.it/> (active 1 March 2007)

⁴⁶ Methods Network, Expert Seminar on History and Archaeology, <http://www.methodsnetwork.ac.uk/redist/pdf/es4programme.pdf> (active 1 March 2007)

⁴⁷ An audio version of his paper is available at: Methods Network, http://www.methodsnetwork.ac.uk/redist/audio/eseminars/es04/es4_01.mp3 (active 1 March 2007)

⁴⁸ University of Portsmouth, Great Britain Historical Geographical Information System, <http://www.port.ac.uk/research/gbhgis/> (active 1 March 2007)

⁴⁹ Vision of Britain, <http://www.visionofbritain.org.uk/index.jsp> (active 1 March 2007)

A similar initiative, also launched in 2001 is the China Historical GIS, hosted at Harvard, which features historical maps and datasets that are freely available for academic use. (See Fig. 5)



Fig. 5 Maps from the CHGIS

The use of GIS as a ‘hub’ technology around which an archive can be built is described in the AHDS Guide to Good Practice: *A Place in History: a guide to using GIS in historical Research*. It quotes an example from the Perseus Digital Library,⁵⁰ which features the Edwin C. Bolles Collection of the History of London. This includes maps of London from different periods presented as different layers, over which place names can be compared and references made to source texts.

Returning to the first project mentioned in this paper, the ‘Valley of the Shadow’,⁵¹ GIS is also a component of the functionality of this project and over 2000 individual dwelling places, taken from a very detailed map of Augusta County, Virginia (1870), are precisely pinpointed for the purposes of allowing users of the system to be able to identify the houses of particular individuals featured elsewhere on the site.

Conclusion

By looking at the various parts of the information cycle throughout this paper, starting perhaps perversely with presentation; going through data structuring and modelling; onto query and analysis; and then back to presentation again, it has become clear that although the selection of tools, processes and resources referred to has been highly selective, they are symptomatic of a great richness of digital techniques that can be exploited by historical researchers. As Greg Krane stated at a recent Methods Network workshop,

Thirty years ago the dog was the print world and the tail was a few ancillary digital tools to augment that. I would say now the dog is the digital world and the print world is the tail.⁵²

In a discipline so dependent on the gathering and analysis of written sources, those doubting the efficacy and authority of digital techniques may find themselves in grave danger of being left at the periphery of historical research.

The references and examples used in this paper include work that has been presented or referred to at a number of Methods Network sponsored events, all of which have some material relating to them on the Methods Network website. In the context of workshops and seminars, a report is produced by the organiser of the event which summarises and draws conclusions from proceedings. Where the event has been

⁵⁰ Tufts University, Perseus Digital Library, <http://www.perseus.tufts.edu/> (1 March 2007)

⁵¹ University of Virginia, The Valley of the Shadow, <http://valley.vcdh.virginia.edu/> (active 1 March 2007)

⁵² Methods Network, Open Source Critical Editions Workshop, <http://www.methodsnetwork.ac.uk/activities/act9.html> (active 1 March 2007)

designated an 'expert seminar', more detailed material such as draft papers, slides or audio recordings of presentations is made available. A list of Methods Network funded events that have broad relevance to historical research can be found at the end of the following 'reference' section.

Neil Grindley

Senior Project Officer

December 2006

(last revision 26 March 2007)

ACKNOWLEDGMENTS

Thanks very much to Craig Bellamy (AHDS), Zoe Bliss (AHDS History), Torsten Reimer (Methods Network) and Matthew Woollard (UK Data Archive) for their assistance in the preparation of this paper.

REFERENCES

PRINT SOURCES

Boonstra, O., Breure, L., Doorn, P. (2004), *Past, Present and Future of Historical Information Science*, NIWI-KNAW,

Denley, P. (1994), 'Models, Sources and Users: Historical Database Design in the 1990's', *History and Computing*, Vol.6, No.1, pp.33-43

Harvey, C., Press, J. (1996), *Databases in Historical Research*, Macmillan, London

Thomas, W. (2004), 'Computing and the Historical Imagination', in Schreibman, S., Siemens, R., Unsworth, J., (eds), *A companion to Digital Humanities*, (pp. 56 - 68)

WEBSITES CONSULTED

Table of Contents
Analysis
Cultural Heritage
Data Mining
Data Structuring
Digital Archives & Libraries
General Comment
GIS
Journal Articles
Organisations
Projects
Quantitative Methods
Query & Retrieval
Resources
Simulation
Standards
Theory
Visualization

Analysis

AHRC ICT Methods Network, Centre for Computing in the Humanities, Kay House, 7 Arundel Street, London, WC2R 3DX.

Topic Detection and Tracking

<http://www.nist.gov/speech/tests/tdt/tdt98/index.htm>

Various tools and reports including automatic speech recognition

TDA

<http://www.stat.ruhr-uni-bochum.de/tda.html>

Freeware for event history analysis

Markov Chain Definition

http://pespmc1.vub.ac.be/ASC/MARKOV_CHAIN.html

Method based on probability analysis

CHNM Tool

<http://www.zotero.org/>

Web site reference and annotation tool

Cultural Heritage

DIGICULT

http://www.digicult.info/pages/pubpop.php?file=http://www.digicult.info/downloads/dc_thematic_issue7.pdf

Cultural Heritage Tools and Methods

Data Mining

Dan Cohen on Data Mining

<http://www.dlib.org/dlib/march06/cohen/03cohen.html>

Refers to syllabus finder and H-Bot

NORA

<http://www.noraproject.org/nora.05.report.pdf>

Major Data Mining Project - Report to the Mellon Foundation (2005)

ARMADILLO

<http://www.hrionline.ac.uk/armadillo/sources.html>

Humanities Research Institute – Sheffield University Data Mining Project

Data Structuring

KLEIO link

<http://rubens.anu.edu.au/htdocs/chart/jaritz.html>

Everyday Life in the Middle Ages and Digital Image Analysis

KLEIO manual

<http://wwwuser.gwdg.de/~mthalle2/manual/tutorial/intro.htm>

Principles and features of KLEIO system

COEL database

<http://ahds.ac.uk/creating/case-studies/coel/index.htm>

Broadly conceived prosopographical database

AHDS bibliography

<http://www.ahds.ac.uk/history/creating/guides/digitising-history/sect72.html>
up to 2000, covers all works on database design

Bruno Boute's paper on Uniform Database Structures
<http://www.linacre.ox.ac.uk/Files/Pros/boute.doc>
Lovanienses Database

Gamera Project
<http://dkc.jhu.edu/gamera/html/overview.html>
Gamera is a framework for the creation of structured document analysis applications by domain experts

SQL and XML
<http://www.sqlx.org/>
Organisation to develop the SQLX standard

VICODI Project
<http://www.vicodi.org/about.htm>
EU project to develop a History ontology

Digital Archives and Libraries

Perseus
<http://www.perseus.tufts.edu/>
Major Digital Library at Tufts

The VOC Archives
http://www.tanap.net/content/voc/organization/organization_intro.htm
Home page of archives relating to the Dutch United East India Company

The National Archives
<http://www.ndad.nationalarchives.gov.uk/>
Searchable datasets pertaining to national statistics

British History Online
<http://www.british-history.ac.uk/>
Digital Library for primary and secondary sources of medieval and modern history of the British Isles

Scholarly Editions, Historian's Archives and Digital Libraries
http://dlist.sir.arizona.edu/1473/01/Dalbello_ASIST_SI.pdf
The pragmatics and the rhetoric of digital humanities scholarship

HISTPOP
<http://ahds.ac.uk/history/about/projects/ohpr.htm>
The Online Historical Population Reports Website: A collection of British Historical Population Reports – funded by the JISC digitisation programme

General Comment

William J Turkel Blog
<http://digitalhistoryhacks.blogspot.com/>
Lots of interesting comment about digital history activities

Digital History

<http://chnm.gmu.edu/digitalhistory/>

Book from Cohen and Rosenzweig covering the creation and use of online history materials

Dan Cohen's Section

http://www.dancohen.org/publications/#digital_history_raw_and_the_cooked

Taken from the publication, Digital History

The Blog for the Carroll R. Pauley Memorial Endowment Symposium

<http://digitalhistory.wordpress.com/>

at the [Department of History](#) at the [University of Nebraska-Lincoln](#)

GIS

GIS in Historical Research

<http://www.ahds.ac.uk/history/creating/guides/gis/index.html>

Ian Gregory's guide to GIS for historians

Great Britain Historical GIS Project

<http://www.port.ac.uk/research/gbhgis/>

University of Portsmouth – Britain's localities as they've changed over time

OxArchDigital

<http://www.oxarchdigital.com/projects.php>

Commercial database and information systems design company

National Atlas Map Maker

<http://www.nationalatlas.gov/>

Map of America in Layers

Journal Articles

Social Science History Article 1

http://muse.jhu.edu/journals/social_science_history/v030/30.2kaufmann.html

Statistics relating to the Orange Order in Scotland - Ecology

Social Science History Article 2

http://muse.jhu.edu/journals/social_science_history/v030/30.3brown.html

Article on Japanese Arable Commons Land – Digital Maps

Organisations

AHC-UK

<http://www.ahc.ac.uk/confweb/2006/conf06.htm>

2006 one day conference

AHC site

<http://grid.let.rug.nl/ahc/>

Association of History and Computing

IATH

<http://www.iath.virginia.edu/iathrails/projects/homepage>

Projects run at IATH in Virginia

National Endowment for the Humanities
<http://www.neh.gov/grants/digitalhumanities.html>
resources for funding and project information

Stoa Consortium
<http://www.stoa.org/?p=488>
News, projects and links for digital classicists

Center for History and New Media
<http://chnm.gmu.edu/>
Develops tools for historians – including automatic online citation tool, web scrapbook, scribe etc ...

The Centre for Contemporary British History
<http://icbh.ac.uk/welcome.html>

University of Massachusetts at Amherst
<http://www.umass.edu/sadri/ssh-journal/>
Social Science History Page

Projects

AHRC funded projects
<http://ahds.ac.uk/history/collections/ahrc-funded.htm>
AHDS History list of AHRB/AHRC Panel 4 funded projects: Medieval and Modern History

List of Resourcing Resources workshops
<http://www.linacre.ox.ac.uk/prosopresources.html#end>
Useful list of projects brought together under the aegis of the Resourcing Resources Workshop

Parliamentary Papers Online
<http://www.bopcris.ac.uk/18c/>
Also JISC funded projects – using advanced technology

Newspapers Digitisation Project
<http://www.bl.uk/collections/britishnewspapers1800to1900.html#strategy>
JISC funded digitisation project

Scholarly Digital Editions
<http://www.sd-editions.com/>
Peter Robinson's output

Virginia Center for Digital History
<http://www.vcdh.virginia.edu/>
Host of the Valley of the Shadow project

Valley of the Shadow
<http://valley.vcdh.virginia.edu/>
Edward L. Ayers and William Thomas' influential and award winning project

Quantitative Methods

New methods for Social History

<http://www.cambridge.org/catalogue/catalogue.asp?isbn=0521655994>

Features articles on a variety of methods borrowed from Social Sciences (1999)

History by Numbers

<http://www.history.ac.uk/ihr/Focus/Whatishistory/books.html>

Review of an Introduction to Quantitative Approaches

Statistics Packages

<http://freestatistics.altervista.org/stat.php>

Open Source and freeware packages

Query and Retrieval

Ontogator

<http://www.seco.tkk.fi/publications/2003/hyvonon-saarela-viljanen-ontogator-combining-view-2003.pdf>

View and ontology based searching with semantic browsing

Dan Cohen article - CHNM

<http://chnm.gmu.edu/resources/essays/essay.php?id=38>

On automatic answering for historical questions – Hbot

H-Bot

<http://chnm.gmu.edu/tools/h-bot/>

Question interface for H-Bot

Mitsubishi Electric Research Labs

<http://www.merl.com/projects/PDH/>

Personal Digital Historian and links to other research projects

Resources

Matthew Woollard on Digital Resources

<http://www.history.ac.uk/digit/woollard.html>

Digital Resources challenging use or users' challenges – points to other references

York University Guide to Resources on the Internet

<http://www.york.ac.uk/teaching/history/pj/pg/internet.htm>

Advice and links to historical resources on the web

Intute Virtual Training Suite

<http://www.vts.intute.ac.uk/he/tutorial/history/>

Internet for historians – online tutorial for web resources

Vision of Britain Through Time

<http://www.visionofbritain.org.uk/index.jsp>

Topographical search by region and location for time-based histories

List of History Resources on the web

<http://www.digital-librarian.com/history.html>

Very large list of all sorts of resources from all periods

ULL – Resources (History)

<http://www.ull.ac.uk/subjects/guides/historytools.shtml>

List of links to web info about History tools and resources

Peer review report

http://www.history.ac.uk/digit/peer/Survey_report2006.pdf

IHR report on peer review and evaluation of digital resources

Peer Review conclusions paper

http://www.history.ac.uk/digit/peer/Peer_review_report2006.pdf

AHRC funded report – as above

IHR list of web resources

http://www.history.ac.uk/guides/internet_historians.pdf

HUMBUL produced list of resources

INTUTE

<http://www.intute.ac.uk/artsandhumanities/cgi-bin/browse.pl?id=200357>

Pages relating to History and Computing

Simulation

Cyberboard

<http://cyberboard.brainiac.com/download.html>

Simulation software

Conflict Simulation

<http://www.kcl.ac.uk/depsta/wsg/consim.html>

King's College War Studies Department

SOCSIM

<http://www.demog.berkeley.edu/~wachter/socstory.html>

Brief history of SOCSIM – a demographic computer microsimulation program (1970)

Hermann and Hermann on the first world war

[http://links.jstor.org/sici?sici=0003-0554\(196706\)61%3A2%3C400%3AAATSTO%3E2.0.CO%3B2-E](http://links.jstor.org/sici?sici=0003-0554(196706)61%3A2%3C400%3AAATSTO%3E2.0.CO%3B2-E)

simulation studies

Standards

Dublin Core

<http://dublincore.org/groups/kernel/>

DC Kernel and ERC (Electronic Resource Citation) working group

CIDOC-CRM

<http://cidoc.ics.forth.gr/>

Provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.

ICOM-CIDOC

<http://www.willpowerinfo.myby.co.uk/cidoc/>

The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC)

Theory

Activity Theory

<http://www.edu.helsinki.fi/activity/pages/chatanddwr/chat/>

Cultural-Historical Activity Theory based on Russian psychology from 1920's

Event History Analysis

<http://faculty.washington.edu/djholman/csss544/index.html>

University of Washington Social Sciences course outline for Event History analysis tools

Multilevel Regression Techniques

<http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html>

Are Multilevel Techniques Necessary? An overview, including Simulation Studies

Multiple Regression

<http://www2.chass.ncsu.edu/garson/pa765/regress.htm>

Very detailed article featuring information about regression techniques

Prosopography at Linacre

<http://www.linacre.ox.ac.uk/prosopo.html>

Outline of prosopographical studies

Review of Cognitive Sci/Text from Linguist

<http://www.ling.ed.ac.uk/linguist/issues/14/14-2171.html>

Chapter by chapter summary of contents of a book about thematics

Visualization

With Digital Maps, Historians Chart a New Way Into the Past

<http://chronicle.com/temp/reprint.php?id=8kffsjkdfxm3yms47t68zbld4pq2vly3>

Article in the Chronicle of Higher Education promoting visualization of historical data

Picturing the Past

<http://www.hpl.hp.com/news/2000/oct-dec/3dimaging.html>

HP labs using angled lighting techniques to discover surface information

VISTA project

<http://www.iaa.bham.ac.uk/bufau/research/wh/Base.html>

Wroxeter Hinterland Project

King's Visualization Lab

<http://www.kvl.cch.kcl.ac.uk/>

3D visualization projects based at King's College London

METHODS NETWORK EVENTS BROADLY RELATED TO HISTORICAL RESEARCH

<http://www.methodsnetwork.ac.uk>

- Virtual History and Archaeology, A Methods Network Expert Seminar on History and Archaeology, Humanities Research Institute, University of Sheffield, 19 – 21 April 2006
- Digital Restoration for Damaged Documents, A Methods Network Workshop, Oxford University Computing Services, Oxford, 29 June 2006
- Historical text Mining, A Methods Network Workshop, Lancaster University, Lancaster, 20-21 July 2006
- Technical Innovation in Art Historical Research, A Methods Network seminar, Centre for Computing in the Humanities, King's College London, 20 November 2006
- Approaches to the Forensic Investigation of Primary Textual Materials, A Methods Network workshop, Humanities Research Institute, University of Sheffield, January 2007
- Theoretical Approaches to Virtual Representations of Past Environments, Goldsmith's College, University of London, 7 March 2007