Past, Present and Future of Historical Information Science

Onno Boonstra, Leen Breure and Peter Doorn[*]

Abstract: This report evaluates the impact of two decades of research within the framework of history and computing, and sets out a research paradigm and research infrastructure for future historical information science. It is good to see that there has been done a lot of historical information research in the past; much of it has been done, however, outside the field of history and computing, and not within a community like the Association for History and Computing. The reason is that the AHC never made a clear statement about what audience to address: historians with an interest in computing, or historical information scientists. As a result, both parties have not been accommodated, and communications with both 'traditional' history and 'information science' have not been established. A proper research program, based on new developments in information science, is proposed, along with an unambiguous scientific research infrastructure.

Chapter 1. Introduction

> "The historian who refuses to use a computer as being unnecessary, ignores vast areas of historical research and will not be taken serious anymore" (Boonstra, Breure and Doorn, 1990).

When we wrote the lines above, fifteen years ago, we sensed that, with the coming of the computer, not only new areas of historical research would be opened, but also that computers would be able to help find solutions to many of the information problems that are so distinctive to historical science.

[*] Address all communications to: Onno Boonstra, Leen Breure and Peter Doorn, NIWI, Postbus 41950, 1009 DD Amsterdam, the Netherlands; e-mail: o.boonstra@let.kun.nl, leen@cs.uu.nl, peter.doorn@niwi.knaw.nl.
The online version of this report is also available on: http://www.niwi.knaw.nl/nl/geschiedenis/onderzoek/onderzoeksprojecten/past_present_future_of_historical_information_science/draft_report/toonplaatje.

Nowadays, information problems in historical research still exist and are still vast and very varied. They range from textual problems (what is the word that is written on this thirteenth-century manuscript? what does it mean? to which issue does it relate? why was it put there? why was the text written? who was the author? who was supposed to read the manuscript? why has it survived?) and linkage problems (is this Lars Erikson, from this register, the same man as the Lars Eriksson, from this other register?), to data structuring problems (how can historical contextual information be kept as metadata in a XML-database?), interpretation problems (from this huge amount of digital records, is it possible to discern patterns that add to our knowledge of history?) and visualisation problems (how do you put time-varying information on a historical map?).

But this does not mean that nothing has been achieved over the last two decades. On the contrary, hundreds of research projects have been initiated to tackle problems like these. Historians, linguists, literary scholars, information scientists, they all have done their share in making historical information science grow and flourish.

Nevertheless, if we look back at what "history and computing" has accomplished, the results are slightly disappointing. 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.

The primary aim of this report is to find out what, when and why things went wrong. A major chapter therefore is dedicated to the Past (Chapter 3), and the way it effected the Present (Chapter 4). In both chapters, attention is focused on *content* as well as *infrastructure*, because both elements – the content of "history and computing" research, and the infrastructure in which this research has been done – have had an impact on the present situation of historical information science.

But disappointment has not been the major incentive to write this report. It is also written to show how much has been accomplished within the field of "history and computing" and what great opportunities lie ahead for further research in computerised methods to be used in historical science.

As a consequence, the report ends with a few suggestions about the future of historical information science. Again, its future is not only a matter of generating new content for historical information science, but also about setting up a new infrastructure. Both issues will be discussed in Chapter 5.

At this point, the concept of "*historical information science*" is introduced instead of "history and computing". This is done deliberately so. "History and computing" is a very vague and confusing term. Historical information science is neither "history" nor "computing". It is a science of its own, with its own methodological framework. The object of historical information science is

historical information, and the various ways to create, design, enrich, edit, retrieve, analyse and present historical information with help of information technology. In this way, historical information can be laid out as a sequential phases of a *"historical information life cycle"*. In Chapter 2, a definition of historical information science is given, as well as a short description of the life cycle of historical information.

The reason for writing a report on past, present and future of historical information science has been that the KNAW, the Royal Netherlands Academy of Sciences, has decided to close down the NIWI, the Netherlands Institute for Scientific Information Services. Part of its replacement will be a new research programme, which will explore the use of information technology in the Humanities. Some of the themes, which could usefully be explored in such a programme, can be discerned by having an insight into what historical information science has – or has not – accomplished over the last years. This background will also allow for the formulation of the infrastructure in which historical information science can thrive.

Normally, there is only limited time and resources for such a report as this; it is therefore highly gratifying that the Department of History of the NIWI has given us the opportunity to study both issues in such depth.

Chapter 2. Historical Information Science

2.1. E-science, e-humanities, e-history

Computing in the sciences and humanities has developed enormously over the past decades. There is hardly a scientist or scholar remaining who does not use a computer for research purposes. There are different terms in use to indicate the fields that are particularly oriented to computing in specific disciplines. In the natural and technical sciences, the term "e-Science" has recently become popular, where the "e" of course stands for "electronic". Although there are now hundreds if not thousands of compositions of "e" *cum* substantive (e-Business, e-Culture, e-Learning, e-Social Science, and even e-Love, e-Death and e-Pizza) in use, the term "e-Humanities" is less common than "humanities computing" and "e-History" is usually called simply "history and computing".

Computer science or informatics as an independent discipline dates to only about 1960. It is often defined as the study of the systematic processing of information by computer systems. In many scientific fields the application of computers has been introduced as an auxiliary or subdiscipline.

Information science is the discipline that deals with the processes of gathering, manipulating, storing, retrieving, classifying and transferring information. It attempts to bring together concepts and methods from various disciplines such as library science, computer science and communication science.

Is it possible to discuss (historical) information science without first defining the term 'information'? Many definitions of the concept exist, but it is not our intention to give an extensive overview here. However, a few words on how 'information' is generally understood and used may be useful. All definitions agree that *information* is something more than *data* and something less than *knowledge*. In the Open Archival Information System (OAIS), information is defined as any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data.[1] According to McCrank, "data are what is given in the smallest units [...] These are the raw material for building information". He gives an admirable list of definitions ('in ascending order'), ranging from data, via information, facts, evidence, and proof, to knowledge, belief, and ultimately wisdom. (McCrank, 2002, 627-628).

In this section, we will go into some conceptual aspects of computing in the natural and technical sciences, in the humanities and in history. Although there appears to be some ambiguity in the usage of terms, we will try to present some common definitions.

2.1.1. E-science and e-Social Science

The term E-Science is predominantly used in Great Britain, where an official and generally used definition appears to exist:

> "science increasingly done through distributed global collaborations enabled by the Internet, using very large data collections, tera-scale computing resources and high performance visualisation."

This definition by the Department of Trade and Industry is supported by the Research Council e-Science Core Programme. Many researchers and institutes, such as the Particle Physics and Astronomy Research Council (PPARC) interpret this definition widely, to include computational and data grid applications, middleware developments and essential hardware procurement.[2] The Oxford e-Science Centre uses the same definition to describe its core activities.

The World Wide Web gave us access to information on Web pages written in html anywhere on the Internet. A much more powerful infrastructure is needed to support e-Science. Besides information stored in Web pages, scientists will need easy access to expensive remote facilities, to computing resources - either as dedicated Teraflop computers or collections of cheap PCs - and to information stored in dedicated databases.

The Grid is an architecture proposed to bring all these issues together and make a reality of such a vision for e-Science. Ian Foster and Carl Kesselman, inventors of the Globus approach to the Grid define the Grid as an enabler for

[1] The OAIS reference model is approved as ISO Standard 14721:2002.
[2] See www.pparc.ac.uk.

Virtual Organisations: "An infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources" (Foster and Kesselman, 1999). It is important to recognize that resources in this context include computational systems and data storage and specialized experimental facilities.[3] The computational grid is the next-generation computing infrastructure to support the growing need for computational based science. This involves utilization of widely distributed computing resources, storage facilities and networks owned by different organisations but used by individuals who are not necessarily members of that organisation.

A descriptive way to explain computational grids is by analogy to the electric power grid. The latter provides us with instant access to power, which we use in many different ways without any thought as to the source of that power. A computational grid is expected to function in a similar manner. The end user will have no knowledge of what resource they are using to process their data and, in some cases, will not know where the data itself came from. Their only interest is in the results they can obtain by using the resource. Today computational grids are being created to provide accessible, dependable, consistent and affordable access to high performance computers, databases and even people across the world. It is anticipated that these new grids will become as influential and pervasive as their electrical counterpart.

In the social sciences, a National Centre for eSocial Science (NCeSS) has recently been formed by the ESRC (Economic and Social Research Council) in the UK. In parallel to the developments in the natural sciences, the NCeSS aims to stimulate the uptake and use by social scientists of new and emerging Grid-enabled computing and data infrastructure, both in quantitative and qualitative research.[4]

2.1.2. E-humanities or humanities computing?

> "Humanities computing is an academic field concerned with the application of computing tools to arts and humanities data or to their use in the creation of these data." (McCarty, 1999)

Humanists have used computers since the 1950s, but until the 1980s usage could be described as occasional (1993). It is clear from the literature and online resources that, especially since the 1980s, computing has pervaded every conceivable field of the humanities, although in some areas the role of computers has become more important than in others.

The introduction of computing in the humanities was not universally met with enthusiasm by researchers. There have been debates on the use of com-

[3] See: http://www.nesc.ac.uk/nesc/define.html.

[4] NCeSS will consist of a co-ordinating Hub and a set of research-based Nodes distributed across the UK. The Hub will be based at the University of Manchester, with support from the UK Data Archive at the University of Essex.

puters in the humanities ever since they were introduced. Even today there are pockets of stubborn resistance against computing. At the same time we can see that, although basic computing skills of word processing, e-mailing and web browsing are nowadays omnipresent among humanities scholars, their methodical and technical skills for computerised research are fairly limited. In 2004 the steep learning curve of such techniques, which was already observed in a report by the Commission for the Humanities of the KNAW in 1997, is as steep as ever.

It is not questioned that the electronic media are extremely important for opening up sources for research that would otherwise remain hidden, inaccessible and impossible to analyse. Digital media are undoubtedly more suitable for source publications than paper, and in many respects also more than microfilm. It is therefore not surprising that many source publishers have turned digital.

Many fields in the humanities are based on the study of documents, handwritten or printed, consisting of text, numbers and images. Other fields are based on oral sources (speech) or on sound (music), on material objects (works of art, archaeological objects, *realia*), or on visual information (photographs, film).

It is a matter of epistemological debate how fundamental the rise of computing (and more in particular of the Internet) is for the ways in which knowledge is produced and used in the humanities. Clearly, the growth of the Web is changing the behaviour and priorities of scholars in a number of respects, but the significance of these changes is only partly understood. Although the importance of the Internet can hardly be overestimated, it is not sure that what we are seeing is a fundamental change rather than the adaptation of what researchers have been doing before they used computers to a new environment.

Nevertheless, information technology has brought changes in certain disciplines of the humanities that can be labelled revolutionary. For example in linguistics, research of language and speech has developed into a sophisticated, thoroughly technological field. In archaeology the application of geographical information systems is now rule rather than exception. Many historical dissertations are literally computer-based, that is: grounded on sources that have first been converted to databases. Text corpora have become indispensable in a lot of literary analysis.

Access to research data was and is a central issue in the humanities. The digitisation of catalogues and other access tools by libraries, archives, museums and other heritage institutions was a first step to improved access to holdings. There is of course a virtually endless amount of potentially relevant material for humanities research, that currently is in analogue form and that could be digitised. Given the variety of disciplines within the humanities and the infinite number of subjects that can be (and are) studied, it is unavoidable that any digitisation programme is selective.

Centres for humanities computing, which are focal points in research, teaching and services on ICT and the humanities, have been created in many countries. Again, we limit ourselves to a few examples. In the UK, mention should be made of the Arts and Humanities Data Service (AHDS), which consists of five specialised centres for a variety of humanities fields distributed over the country.[5] The Office for Humanities Communication at King's College, London, is an umbrella organisation that fosters communication among scholars and others involved in computer-related projects and activities. It has published a series of monographs and collected papers concerned with the impact of computers in humanities scholarship and higher education.[6] The Humanities Computing Unit at Oxford carries out research and develops resources in many areas of humanities computing. The unit provides support for academics in the humanities applying new technologies to their research and teaching. Among other things, the unit includes the Centre for Humanities Computing, the Oxford Text Archive, and the Humbul Humanities Hub.[7] The mission of the Humanities Advanced Technology and Information Institute (HATII) at Glasgow University is to actively encourage the use of information technology and information to improve research and teaching in the arts and the humanities.[8] Recently, the Arts and Humanities Research Board has announced the establishment of an ICT Methods Network. The aim of this network is to promote and disseminate the use of ICT in UK Arts and Humanities research in order to enhance, develop and make more effective the process of research, and to communicate research outcomes more widely and efficiently. It will focus on new developments and advanced methodologies, on research processes, questions, and methods, and on uses of data, rather than on data creation and preservation or access to resources.[9]

In the USA, we limit ourselves to mentioning the Centre for Electronic Text in the Humanities at Princeton and Rutgers Universities and the Institute for Advanced Technology in the Humanities at the University of Virginia.[10] In Canada, the mandate of the Computing in the Humanities and Social Sciences facility at the University of Toronto is to promote computing in research and teaching within the humanities and social sciences.[11] The list could continue for many pages with references to centres and institutes in humanities computing in many European countries and on other continents. The developments within

---

[5] See: http://ahds.ac.uk/.

[6] http://www.kcl.ac.uk/humanities/cch/ohc/index.html.

[7] http://www.hcu.ox.ac.uk/index.html; Humbul is an excellent portal for humanities computing. The Oxford Text Archive hosts the AHDS Language, Literature & Linguistics: http://ota.ahds.ac.uk/.

[8] http://www.hatii.arts.gla.ac.uk/.

[9] http://www.ahrb.ac.uk/apply/research/ICT/ahrb_ict_methods_network.asp.

[10] http://www.ceth.rutgers.edu/index.htm and http://jefferson.village.virginia.edu/home.html.

[11] http://www.chass.utoronto.ca/.

the Netherlands were comparable to those abroad, and in some fields Dutch 'humanities computing' was comparatively strong.

The journal *Computers and the Humanities*, which was founded as a newsletter as early as 1966, is a good source to get an insight in the broad field of humanities computing. There are also many journals on this subject in the various humanities disciplines, such as *History and Computing*. A series of monographs in the British Library Research series give an excellent overview of humanities computing until the mid-1990s (Katzen, 1990; Kenna and Ross, 1995; Mullings, 1996). A book edited by Terry Coppock (Coppock, 1999), the chairman of the British Academy, extends the picture until the end of the last century. For more recent additions and reflections on the use of ICT in the humanities, some articles, conference papers and, of course, Internet sources are relevant. For instance, the electronic publication *Humanist* is an international electronic seminar on the application of computers to the humanities. It provides a forum for discussion of intellectual, scholarly, pedagogical, and social issues and for exchange of information among members.[12]

Humanities computing is interdisciplinary by definition. The first thing that is striking when we look at definitions of "humanities computing" is that humanities scholars tend to spend a lot of words on the concept, and that they approach the subject with great care, if not with many detours. They even tend to discuss at great length whether Humanities Computing is an academic field or not (see the conference papers from a seminar held at the University of Virginia in December 1999).[13] Geoffrey Rockwell describes his quest about this question as a Table of Digressions:

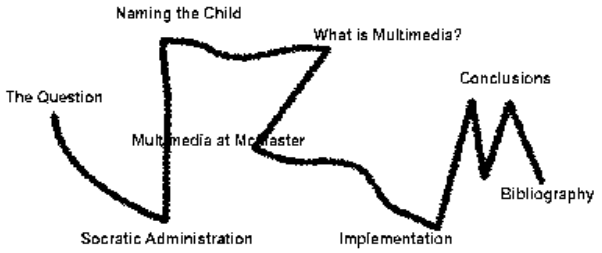


Figure 2.1. Is humanities computing an academic discipline? (Rockwell, 1999)

[12] *Humanist* is allied with the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing. It is an affiliated publication of the American Council of Learned Societies and a publication of the Office for Humanities Communication (UK). See http://www.kcl.ac.uk/humanities/cch/humanist/.

[13] "Is Humanities Computing an Academic Discipline?" *The Institute for Advanced Technology in the Humanities, The University of Virginia*. http://www.iath.virginia.edu/hcs/.

In a terminological paper, Willard McCarty describes the historical development of the term "humanities computing", which was preceded by *computing* **and** *the humanities* (as in the name of the professional journal) and by *computing* **in** *the humanities*. He also considers *humanistic informatics*, but instead of arriving at a firm conclusion he leaves us with two questions:

> "The basic question [...] however, is independent of its idiosyncrasies, namely, what kind of a practice do we want? What approach to the other humanistic practices is best for it?" (McCarty, 1999).

In Susan Hockey's view, the core of the research agenda of humanities computing is knowledge representation and manipulation, or, to use a term that has a broader application in the world beyond academia, it is about information management. In the humanities we are dealing with a variety of extremely complex information (Hockey, 1999). This information can have a very long life span and so electronic representations of it must also last for a long time, far longer than the computer system on which they were created. The information can also be studied for many different purposes and we therefore need to research and develop computer tools not just to deliver the information to people but also to help them study and manipulate the information in different ways. We need also to find out how effective these tools are and to feed the results of evaluations of the tools into new and better systems.

Espen Aarseth tries to find a good balance between what goes inside and outside of the field of humanistic informatics by defining an upper and a lower threshold of the field (Aarseth, 1998). He visualizes this as a pyramid of three levels, where the top represents each traditional departmental field and its specific methodologies. The middle level comprises the activities and applications that do not exclusively belong to any specific field, but are relevant to several. This is where humanistic informatics is located in Aarseth's view. Beneath the lower threshold we find the applications, method and perspectives that do not stand in any particular relationship with humanities research or its research objects.

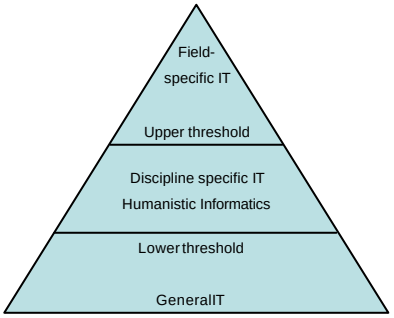


Figure 2.2. Aarseth's three levels of information technology (Aarseth, 1998)

The Department of Humanistic Informatics at the University of Bergen keeps a focus on the methodological tradition, while developing the fields of digital culture analysis and the exploration of new media technologies, with the following directions as key areas:

1) Humanistic IT-methods. Studies of how humanities research applies new digital methods to solve problems within the various disciplines. Examples of this are data analysis by explorative (and traditional) statistics, systems for machine assisted translation, text corpus, dictionaries, database applications (such as lexicography, terminology), tagging and mark-up, geographical information systems, use of simulation and dynamic models in the study of cultural processes, three-dimensional graphical presentation of objects and phenomena.
2) Multimedia and hypermedia research. Understanding and development of multimedia-applications; distributed multimedia platforms and network communication, WWW-programming, hypertext-development and research on standards such as XML, VRML, HYTime, etc.
3) Pedagogical software and the development and use of network communication for pedagogical purposes, such as distance learning. Information- and communication technology (ICT) -based didactics.
4) Digital culture and digital rhetoric and aesthetics. The study of digital modes of communication and topics like computer art, digital literature, Internet cultures, virtual reality, computer games, gender/identity and ICT, through cultural and communication theories.

By combining explorative, methodological, and critical approaches, the Department of Humanistic Informatics aims to offer a uniquely humanistic perspective on ICT, and to create an independent theoretical framework for the study of the new digital fields.

2.1.3. E-history, 'history and computing' and 'historical information science'

> "History as science is an information science. Its specialty is informing the present about the past" (McCrank, 2002)

The discussions and developments concerning the definition of 'history and computing' can be envisaged as a particular subset of (or parallel to) those in 'humanities computing'. Also here, there are different names for the field in use. The names and definitions used are partly dependent on ideas about the field and partly dependent on language. In Dutch, apart from '*geschiedenis en informatica'* the term '*historische informatiekunde*' is common. The latter term refers to applied informatics and information science in the historical discipline. Most historians regard computing in historical research as a technical and auxiliary trade. In English, 'history and computing' is the most neutral and encompassing term, while 'historical information science' refers to the specific

field of history in the more general discipline of information science. 'Cliometrics' is oriented on historical econometrics (or quantitative economic history; Clio being the muse of history). In German the term *'historische Fachinformatik'* is in use, while in Russia '*istoricheskaya informatika'* is used to indicate the field. Also the term 'historical information processing' is used internationally.

In a recent survey of the literature, Lawrence McCrank proposes to define historical information science as a hybrid field of study:

> "Historical information science integrates equally the subject matter of a historical field of investigation, quantified social science and linguistic research methodologies, computer science and technology, and information science, which is focused on historical information sources, structures, and communications." (McCrank, 2002)

In contrast to this, according to Charles Harvey, historical computing must be concerned with the creation of models of the past or representations of past realities. It cannot be defined simply in terms of areas of application or applied information technology. Database systems or expert systems might happen to be of tremendous interest, but there is nothing specifically historical about such things. They are just general tools. Historical computing can only be defined in terms of the distinctive contribution it can make to historical research. As a subject, it exists on the methodological plane, and none of its historical methods owes anything to computers as such: *historical computing can be done without computers*. Computers merely make operational the concepts and methods that are the product of historical computing. Historical computing is a formal approach to research, that requires data and algorithms to be made explicit, and, as such, it is part of scientific history[14].

Historical Informatics has been defined as a new field of interdisciplinary specialization dealing with pragmatic and conceptual issues related to the use of information and communications technologies in the teaching, research and public communication of history.[15] It reflects the reality that the history-based disciplines are being transformed through the impact of new technologies, to the point where new interdisciplinary norms and practices have emerged. As such, Historical Informatics recognizes that the traditional division of labour between professional historians and information specialists fails to meet the needs of either group.

To better cope with the demands of practitioners in the digital publishing era, historians need to become more information literate, while information specialists need to better understand and the specific information needs of historians. Arguably, what is needed is a new generation of practitioners who

---

[14] Italics are all ours.

[15] Welling has proposed to use the term 'Computational History' in an analogy of 'Computational Linguistics' (Welling, 1998). However, computational history is often understood as the history *of* computing.

are highly trained in the craft of history as well advanced information skills, such as computer programming, database development and multimedia production.

However, history and computing is not only about historical research, but also about historical resource creation (Woollard, 1999). More and more archival sources are becoming available in digital form. On the one hand, researchers are transforming archival sources into digital files like modern monks (Doorn, 2000). Part of this material is made available for secondary analysis by other researchers through data archives. Reasoned and commented source editions, which in the past were only published as books, are now increasingly being published in electronic form. Archives and other heritage institutes are also digitising parts of their collections. Presently, only a fraction of the archival holdings are available in digital form, but the amount is constantly growing.

The growing availability of digital historical sources is bringing about a change in the set-up and organisation of historical research. By bringing sources virtually together that are physically stored scattered in archives around the globe (such as the VOC archives in Europe, Asia and Africa), new opportunities for comparative research emerge, that were unfeasible in the past. Digital source collections can be studied from different perspectives by larger groups of researchers in the form of virtual 'collaboratories' on the Web. Others can also check research if the digitised sources are made available.

Another area in which new research possibilities are being created by the growing availability of vast quantities of digital material, is that of electronic image repositories (of photographs, paintings, etc.). Historians traditionally use pictures as illustrative material, but now systematically unlocked historical image collections can be used for direct analysis. In addition to this, there is an exponentially growing modern archive of data banks, e-mail correspondences, and specialised information systems that only exist in digital form (Higgs, 1998). The historical research of the future will increasingly depend on the processing of this 'digitally born' material. Of course, there is a whole range of issues related to the problems of 'digital longevity', many of which require further research.[16]

In this report, we propose to define historical information science (*historische informatiekunde*) as the discipline that deals with specific information problems in historical research and in the sources that are used for historical research, and tries to solve these information problems in a generic way with the help of computing tools.

This definition not only excludes specific information problems outside the field of historical research, but also general information problems that are not

[16] There is a fast growing body of literature on digital longevity, to which archivists, information scientists and digital librarians contribute. Also in some commercial sectors the digital preservation of documents and research materials plays an important role, e.g., in the pharmaceutical sector.

specific for historical research. Our definition of historical information science is compatible with the Aspen's definition of humanistic informatics.

It is important to dwell upon the clause "in a generic way" in the above definition. Historical information science is not about finding specific solutions for specific historical problems, but about finding solutions that transcend the specific problem. To reach that goal, specific problems need to be formalised or generalised, before more generic solutions can be found. Of course, the more general solution in the form of a tool or demonstrator will also be applicable in the specific case.

So, "historical information science" is about determining and solving information problems. The information problems in historical research can be divided into four categories:

– *information problems of historical sources*

Sources are the basis of historical research. They have their own particularities: they contain inconsistencies, they are incomplete or incompletely preserved, and they are unclear or ambiguous. Another characteristic of sources is that they may contain information about different units of analysis: a population register contains information about individuals, about the families to which they belong, and the households in which they live. This type of multi-level problem requires further investigation.

Moreover, not all historians have a clear research question at the start of their investigation: it is therefore impossible and undesirable to model an information system on the basis of an information requirement that has not yet crystallised.

Finally, historical research aims to unravel *"wie es damals gewesen ist"*. This means that the meaning of a certain piece of data cannot exist without interpretations. These interpretations therefore need to be added to the information system that has been built on the basis of the source. However, interpretations are by definition subjective. They need to be added, but in such a way that they can be separated from the original data in the source.

– *information problems of relationships between sources*

When several sources are used in historical research additional problems will arise. In the first place there will be the linking of the data: how can we establish whether a person in source A is the same as another person mentioned in source B? This is particularly problematic if there are spelling variations or more persons with the same name (more on nominal record linkage, including literature references, in section 3.5.1.6). In the second place there is the problem of changes over time and place. The meaning of data from a source is dependent on the spatial and temporal context. When linking such a source to another one, potential changes in meaning need to be considered. For instance, an occupational title such as "labourer" can mean "agricultural worker" at one point in time and "factory worker" at another point in time.

– *information problems in historical analysis*

Historical research deals with changes in time and space. Therefore, historians need analysis tools that take these changes into account. Many statistical analysis techniques that historians use are borrowed from the social sciences. The social-scientific analysis methods that measure changes over time are only suitable in an experimental setting. The other techniques are at best usable for cross-sectional analysis, but not for diachronic analysis. Several techniques have been developed that seem useful for historical research (such as event history analysis), but the applicability of these techniques still needs to be determined.

The same is true for techniques that process information from different units and levels of aggregation. Multilevel regression techniques seem appropriate. However, applications of ecological inference techniques in historical research are yet to be explored.

– *information problems of the presentation of sources or analysis*

In accordance with the lack of specific historical analysis tools, presentational techniques for historical data also require research. For instance, methods to represent changes in time and space, to visualise multi-level linkages, etc.

Historical information problems are different at various stages of historical information. Therefore, an important classification scheme is the life cycle of historical information.

2.2. The life cycle of historical information

Like many objects that take part in a process of production and consumption historical information will go through several distinct stages, each related to a specific transformation, which produces a desired quality. Each of them represents a major activity in the historical research process. Together the stages are referred to as a life cycle. The idea of an information life cycle is derived from records management (sometimes referred to as 'document control'), where the idea of document life cycle is central to the overall process: design and creation of records, approval, retrieval, circulation, access and use, archiving and destruction.

Historical *information* is to be distinguished from raw data in historical sources. These data are selected, edited, described, reorganized and published in some form, before they become part of the historian's body of scientific knowledge. Information once presented may give cause for the creation of other information, in this way closing the circle. The stages of the life cycle are not always passed through in a strict sequence, and some of them may even be skipped under certain circumstances.

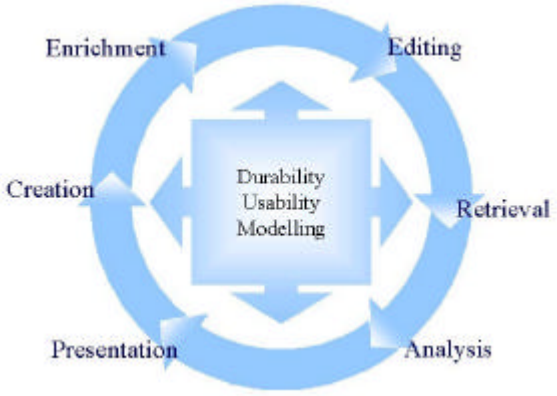


Figure 2.3. The life cycle of historical information

The edge of the life cycle shows six stages:

1) **Creation**. Creation comprises not only the physical production of digital data, but also the design of information structure, (e.g., through data modelling, text modelling), and, in a broader sense, the project's design as well. The practical aspects regard data entry and text entry tools, digitization (like OCR) and considering appropriate database software.
2) **Enrichment**. The bare data, images and texts once created will need to be enriched with metadata, which describes the historical information itself in more detail, preferably in a standardized way (e.g., by means of a system as Dublin Core), and intelligible to (retrieval) software. It comprises also to the linkage of individual data that belongs together in the historical reality, because these data refer to the same person, place or event (so-called nominal record linkage).
3) **Editing**. Enrichment changes into editing, which includes the actual encoding of textual information by inserting mark-up tags, or by entering data in the fields of database records. Enhancement may be considered as a separate phase of the editing process. It is the process by which data is transformed through algorithmic processes preliminary to analysis. Editing extends also to annotating original data with background information, bibliographical references, and links to related passages.
4) **Retrieval**. When information has gone through the previous stages, it is ready to be selected, looked up and used (i.e. retrieval). In some projects this will happen only after a formal publication, which moves retrieval to a later point in the life cycle. Retrieval itself is based on mechanisms of selection and look-up, like SQL-queries in the traditional database environment, and Xpath and Xquery expressions for texts in XML-format. In addition it pertains to the user interface, for

both the specification of the query itself, and for the display of results (in the form of a simple list or through in a more advanced visualized representation).

5) **Analysis**. Analyzing information means quite different things in historical research. It varies from qualitative comparison and assessment of query results, to advanced statistical analysis of data sets.
6) **Presentation**. Historical information is to be communicated in different circumstances through various forms of presentation. Although conceptually represented here as the final stage, it will actually happen frequently in other stages as well. Presentation of digital historical information may take quite different shapes, varying from electronic text editions, online databases, virtual exhibitions to small-scale visualizations of research results within a single project.

In addition, three practical aspects have been grouped in the middle of the life cycle, which are central to computing in the humanities and in different ways related to the six aforementioned stages:

- **durability**, which has to guarantee the long term deployment of the thus produced historical information;
- **usability**, which regards the ease of use, efficiency and satisfaction experienced by the intended audience using the information produced;

and, finally,

- **modelling** in a broader sense than the data modelling or text modelling, mentioned above. Here, modelling refers, amongst other things, to the more general modelling of research processes and historical information systems.

Chapter 3. The Past

3.1. The beginning of History and Computing

The beginning of applying computer techniques to historical material is often associated with the well-known impressive literary and lexicographic projects from the early days of humanities computing. Classic milestones are the production of the *Index Thomisticus* by father Roberto Busa[17] (finally resulting

[17] Roberto Busa was born in Vicenza on November 28th 1913. He started his career as a Jesuit scholar and became a full professor of Ontology, Theodicy and Scientific Methodology. In 1946 he planned the *Index Thomisticus*. The work was mainly carried out in Gallarate and in Milan until 1967, in Pisa till 1969, in Boulder (Colorado) till 1971 and, for the next nine years, in Venice, where, from 1974 till 1980, the photocomposition of the 70,000 pages forming the 56 encyclopaedic volumes of the *Index Thomisticus* was accomplished using IBM computers:
http://www.kcl.ac.uk/humanities/cch/allc/refdocs/honmems.htm#busa.

into 56 printed volumes with analysis of the work of Thomas Aquinas), the extensive electronic editions of Christian Latin texts by the CETEDOC in Louvain from the 1970s onwards[18], and the founding of the widely accepted professional journal *Computers and the Humanities* by Joseph Raben in 1966 (Gilmour-Bryson, 1987). One of the early larger undertakings was *The Philadelphia History Project* (Hershberg, 1981). In art history, Laura Corti directed attention to the relevance of computers (Corti, 1984a, 1984b; Corti and Schmitt, 1984). All this may be regarded as the beginning of history and computing in a broader sense, but history and computing in a more narrow sense, explicitly related to history as a discipline, took shape later. In this respect, the developments in the United States differed from those in Western and Eastern Europe.

In the United States, computer-aided historical research took off in the 1960s, mainly propelled by enthusiasm for social and economic history. Using methods and techniques originating from the social and economic sciences, it focused, therefore, on computation and quantitative aspects (Greenstein, 1997). Both Themstrom's *Other Bostonians* (Thernstrom, 1973) and Fogel and Engerman's *Time on the Cross* (Fogel and Engerman, 1974) are considered milestones in this respect. It did not take long, however, before quantitative history started to find opponents. Cliometrics was severely criticised for placing too much trust in counting and in fragmentating the historical image of the past.

In various countries in Western Europe, social history failed to promote broader interest in the application of computing techniques. In France, for instance, the detailed interest in the social and economic basis of life, so well expressed by the *Annales* group, hardly impacted on historical computing. The first historians to use computers for research purposes, were historical demographers. The Cambridge Group for the study of population and social structure (established in 1964) centred around Peter Laslett led the way.

It may very well be that the traditional character of the history curricula in the European arts faculties did not foster close co-operation with the social sciences, as had happened in the United States. Looking for guidance in computer applications, scholars in the humanities had to rely on help mainly from computer linguists. A great deal of the activity therefore centred around source editing, e.g. concordances and source editions by the CETEDOC in Belgium and the CNRS in France[19], where, amongst others, Lucie Fossier, Caroline Bourlet, and Jean-Philippe Genet from the Institut de Recherche et d'Histoire des Textes, have shaped computer-aided research projects in medieval history,

[18] CETEDOC: CEntre de Traitement Électronique de DOCuments, directed by P. Tombeur. For a list of publications refer to: http://zeus.fltr.ucl.ac.be/recherche/publications/pub_source.html.

[19] CNRS: Centre National de la Recherche Scientifique, in this context particularly relevant its laboratory of texts: L'Institut de Recherche et d'Histoire des Textes (IRHT).

with the bulletin *Le Médiéviste et l'Ordinateur* (founded in 1979) as an important channel for scholarly communication.

In Western Europe, Germany was the exception to the rule. In 1975, German historians and sociologists founded the *Quantum*-group in order to explore, in close collaboration, possibilities and problems with the use of historical and process-produced data.[20] It was driven by a feeling of uneasiness in empirical sciences with data based on surveys only, and by the turn of historians away from ideographic and narrative approaches. It was aimed at closing the gap between the German situation and the upswing of quantitative history elsewhere. A few years later, its journal, *Quantum Information*, changed its name into *Historical Social Research* (*Historische Sozialforschung*) and grew into a broader platform for publication of subjects concerning history and computing, with its focus, however, on the computational aspects of historical research.

At that time, in Eastern Europe, and especially in the former USSR, a remarkable co-operation existed between historians and computer specialists with a background in mathematics. The early experiences with historical computing were related to the processing of statistical data in the field of social and economic history. In the 1970s quantitative history gained a firm footing there, with special interests in problems of historical simulation (Borodkin, 1996).

3.2. Take off: Manfred Thaller's CLIO

In these early days, historians had to rely on general purpose software, available in local computer centres, or had to develop – mostly with some technical assistance – their own more specific tools, which usually boiled down to ad-hoc solutions. Some historians experimented with standard database software: user-unfriendly hierarchical database management systems which required programming for navigation and data look-up, or the not yet fully-grown relational database programs, already popping up sparsely on university mainframes. Once the problems of data storage and retrieval had been overcome, data had to be reformatted for statistical analysis, which required some additional programming.

The typically individual nature of most of these projects defied the potential progress in historical computing that could have been made due to the widespread awakening interest in the prospects of this new technology. This was well understood by the Max Planck Instut für Geschichte in Göttingen, where Manfred Thaller took the initiative in producing dedicated historical software. In 1980, he announced the birth of CLIO (later rechristened to ??e??), a genuine historical database management system (Thaller, 1980). It stood out from other database management systems, among other reasons, by its flexibility in

[20] *Quantum Information* 2 (1977) p. 1-2.

data input formats which reflected the structure of the historical source rather than the requirements of the computer program, a Latin-based command language, but above all by a clear vision on what historical information processing should be. This vision was to be translated into the software's capability to support the process of historical research and the variety of historical sources, without forcing data to be squeezed into standard formats. Moreover, it had some built-in awareness of the fuzziness of historical data and the complex process of historical interpretation, which should prevent hasty conclusions in the data entry stage.

Although CLIO was primarily intended to support ongoing research on the Max Planck Institute itself (which had some emphasis on social and economic history), Thaller offered the CLIO package to research projects elsewhere and looked for co-operation in further software development. Although this offer was necessarily limited to a small number of projects, it heralded a new period of historical computing.

3.3. Getting organised

Apart from CLIO, which was set up to be of use to various historical research projects, historians became aware that it would be worthwhile to share the computational problems they encountered, and the solutions they reached, with their colleagues. Organisations such as the Association for History and Computing were set up in order to facilitate them. In the United States, for instance, the American Historical Association created such a platform for quantitative historians (Greenstein, 1997).

3.3.1. The Association for History and Computing

European historians missed an independent organisation for sharing experiences and establishing an understanding of historical computing. In 1983, an international group of enthusiastic computer using historians gathered for a symposium at Hull. Ambitious ideas circulated, like the launching of an MA in historical computation. The most lasting outcome of this meeting was the *Association for History and Computing* (AHC), established three years later by an inaugural conference at Westfield College (London). In the following decades, international conferences were organised under the auspices of the AHC, from all over Europe to Moscow and Montreal, and professional communication was sustained through its journal *History and Computing*.

The AHC was established at the same time as an enhanced awareness of the pervasive power of computer technology in society was developing. Governments across Europe reacted to this through stimulation programs targeted at the acquisition of basic computer skills by all students. The Dutch government, for instance, decided to stimulate the use of computers particularly in the arts

disciplines and the social sciences. In 1985, *perestroika* favoured the start of a campaign for computerisation in the USSR, which brought the requirement of IT-certificates for university teachers and obligatory computer courses for all students (Borodkin, 1996). However, for a long time wide-scale progress was hindered by a lack of sufficient hardware.

3.3.2. As an example: alfa-informatica in the Netherlands

As a result of additional funding from the mid-1980s onwards Dutch universities enlisted special computer literate staff to teach computer classes directed towards the professional interests of arts students (the so-called *alfa-informatica*) and, to a certain extent, to give support to fellow researchers. This lofty initiative was lacking in one fundamental respect: it did not provide any research infrastructure for historical computing itself. Although most of the staff members had some research opportunities alongside their teaching, research in technical aspects of historical computing remained isolated. Within the framework of NWO (the Dutch agency for research funding) applications for methodological and technical research projects tended to fall between two stools. They were not sufficiently technical to arouse interest amongst computer scientists, and tended to be outside of the intellectual sphere of leading historians then unfamiliar with the use of computers.

Despite this there were further practical organisational changes in the field of computer-aided historical research. In 1987 the Dutch-Flemish/Belgian counterpart of the AHC was founded – the *Vereniging voor Geschiedenis en Informatica* and was soon affiliated with the AHC, followed by the creation of the Dutch Historical Data Archive (*Nederlands Historisch Data Archief* – NHDA) in 1988. This ambiguous situation in the Netherlands, a new domain of academic knowledge without an adequate research infrastructure, reflected the diversity of views on historical computing held in general. Even within the AHC different opinions on the relationship between 'history' and 'computing' managed to live side by side, while 'history and computing' as a particular form of methodological research was only a high priority for some individuals.

3.4. Ideas, claims and convictions

Reflection on the phenomenon of historical computing itself, i.e., the interplay of leading ideas, claims and convictions about suitable methodology, software and the relation of this field to history as a discipline, is far more relevant for an understanding of its past than a reflection on the detailed historiography as unfolded through numerous project papers in the proceedings of the several conferences, journals and workshop volumes. These are, in any case already well documented by (Denley, 1994; McCrank, 2002; Speck, 1994; Woollard,

1999). Consequently, this historical overview will be curtailed and the remainder of this chapter will focus on more philosophical aspects.

The prevailing ideas about historical computing can be reduced to two basic points of view:

- *Plain IT*: This category of view is characterised by a high assessment of the native capabilities of information technology for the historical discipline. Those who support this point of view, tend to praise the possibilities of computer technology as it is (thus 'plain' IT). They will recommend computer usage and the mastering of necessary computer skills. The underlying tacit assumption seems to be that IT-as-available is good enough and covers most, if not all historical requirements; it needs only to be learned and to be applied. If technology fails in certain respects, its shortcomings have to be accepted or to be circumvented. For various reasons, enhancing and adapting the technology itself seem to be beyond their range.
- *Enhanced IT*: Views of this type tend to emphasise the special and complex nature of historical data processing in contrast with computer applications in, for example, business and hard sciences. They show *less confidence in standard information technology* and pay more attention to dedicated software, to special tools, to the implementation of additional knowledge layers, and to fine-tuned methodologies and techniques. Here, the assumption is rather the opposite of Plain IT: information technology as it comes to us should be adapted and extended before it can meet the requirements of historical research – and we should go for it!

It is understood that both points of view are not necessarily contradictory, and allow ample room for different shades of meaning and wise, mixed positions. However, both approaches differ enormously with respect to the direction of historical computing as a field of study. The first one directs attention more to the practice of historical research, representing the computer as a ready-made tool, like a car, a typewriter or a video camera and rates historical computing as hard core historical research. The second calls for investments in dedicated software and favours the development of historical computing into the direction of a historical information science with a theoretical basis.

The AHC conferences in the late 1980s (in particular, Westfield I and II, Glasgow and Cologne) produced fruitful discussions on these subjects, but, unfortunately, did not reached firm conclusions. The later conferences, after 1990, show a predominance of reports on the application of information technology with an emphasis on historical aspects, and a decline in debates on philosophical issues.[21]

[21] Exceptions were the Nijmegen (1994), Moscow (1996) and Tromsø (2003) conferences.

3.4.1. Plain IT

3.4.1.1. Optimism about technological achievements

The first Westfield conference (1986) praised the progress in computer technology and its salutary effects for the historical field. It was impressed by the 'microcomputer revolution' that had taken off. Ann Gilmour-Bryson remembered in her opening paper:

> "Only six years ago, we were forced to learn dBASE programming and to deal with its truly dreadful manual if we wanted to manipulate historical databases on out tiny 16, 32, or 64 K CP/M based early micros. As I typed this article on my IBM PC AT with its 1.2 megabyte floppy disk, its 20 megabyte hard disk, making automatic backup on its 20 megabyte tape backup system, and printed it on an 8 page per minute printer, it was hard to believe that such progress had taken place in six years. Readily available software with a type of artificial intelligence interface enables the novice to build a database, query it, and report from it within a few hours at most."[22]

She concluded that we could now easily test so many hypotheses by simply asking questions of our databases and solve problems that would not have been solved "before the present day because of a lack of time and/or manpower". The computer has become the historian's intelligent and high-speed slave.

In their introduction to the conference proceedings Deian Hopkin and Peter Denley noted that historical computing had evolved considerably from the early days of quantification and basic data processing and now provided a "common ground between historians who otherwise inhabit segmented and secluded worlds". Others praised "the variety of historical uses to which the computer can be put", unburdening historians from tedious, repetitive tasks (Woods, 1987). What drew people together was the enthusiasm of working with large volumes of complex data, now made possible by the blessing of cheap and generally available computer power (Harvey, 1990).

Although this idea of the 'unifying power' of the computer, bringing people together to share their experiences, was a common feeling among the conference participants thanks to the warm hospitality of Westfield College, it was limited in practice and a bit naïve. It did not make clear what, exactly, were the common assets, except, perhaps, for co-operation in the acquis ition of funding. Everybody who used the opportunities to learn more about historical computing, would discover very soon that he had to made choices, particularly with regard to methodology and software to be used, and was thus confronted with the different directions of thought.

3.4.1.2. Extending historical craftsmanship

Of course, from the beginning everybody agreed on the principle that users should learn more about computers in order to use them productively. But this

[22] Gilmour-Bryson (1987), p. 7.

could mean quite different things. Historians, for example, had to learn to be more explicit in their research strategies because computers would require detailed instructions. Some considered this as adding electronic aids to the critical intellectual skill historians had developed since the time of Herodotus and suggested that historians should learn programming as well (Woods, 1987). From this perspective, historical computing had (or at least should have) become entirely part of the historical craft, which had to be modernised due to advances in technology. As Adam Hodgkin expressed it in 1986, speculating about the situation ten years later:

> "It is arguable that there will be no longer a need for a conference on history and computing. By then the uses of computers in historical research will be so well understood and so much a part of the fabric of scholarship that it would be as unnecessary as having a conference on libraries and history. I predict that there will be a history and computing conference in 1996, but I have some sympathy for the view that there is nothing of importance, apart from historical content, that is unique about historical research and computing. *There can be very few computing techniques which are solely of interest to historians*."[23]

The computer is a tool, which, like many other tools, has some general utility in the study of history (Greenstein, 1989). The same idea was expressed in the introduction to the papers of the Glasgow Conferences:

> "*Computing does not have to be complicated*; indeed there is a danger that those who insist that it is so are losing focus of their initial historical enquiry, replacing it with a technology dependent methodology."[24]

At the end of that conference, Charles Harvey philosophised about the nature of historical computing. Looking backward, he expressed ideas that proved to be widespread among computing historians and which have not particularly favoured the growth of historical information science as a methodological discipline. They marshalled feelings and attitudes that justified a turn away from the technical aspects.

Pre-eminently, according to Harvey, historical computing must be concerned with the creation of models of the past or representations of past realities. It cannot be defined simply in terms of areas of application or applied information technology. Database systems or expert systems might happen to be of tremendous interest, but there is nothing specifically historical about such things. They are just general tools. Historical computing can only be defined in terms of the distinctive contribution it can make to historical research. As a subject, it exists on the methodological plane, and none of its historical methods owes anything to computers as such: *historical computing can be done without computers*. Computers merely make operational the concepts and methods that are the product of historical computing. Historical computing is a

[23] Hodgkin (1987), p. 256.
[24] Mawdsley, Morgan, Richmond et al. (1990), p. xi.

formal approach to research, that requires data and algorithms to be made explicit, and, as such, it is part of scientific history.[25]

3.4.1.3. Relying on standards: the relational database.

Others allowed more room for 'foreign imports' from computer science, like a thorough requirements analysis, which would help to determine the feasibility of historical computer projects and could deliver its justification (Greenhalgh, 1987). From the trust in the current state of technology followed naturally the reliance on universal standards, like the relational model for data management.

The application of the relational database model to historical data collections has been defended from the early beginning, e.g. (Greenhalgh, 1987). In the AHC community, this claim has managed to co-exist with arguments for a special historical data model as that of ??e??. Although ??e?? has kept a niche, about the end of this period the hegemony of the relational database was virtually unquestioned among British historians and as scarcely less influential elsewhere (Denley, 1994b). Greenstein looked upon the relational database as a tool particularly suitable for source-oriented data processing. A source-oriented approach should allow for two basic requirements: the same source is handled differently in various stages of historical research and the uses of sources vary over time. A relational DBMS catered very well for the dialectic interpretative process with its resort to the original, because raw source fragments could be copied to database records without any textual change and be linked to standardised data afterwards, allowing thus efficient comparison and analysis while the original text was kept as well (Greenstein, 1989).

At the end of this period Microsoft introduced its popular relational desktop DBMS Access, a wonderful big lie, with respect to the complexities of database design. It was wonderful because of its user-friendly interface. It rapidly swept away its stubborn predecessors like dBASE and Paradox. If a historical data set was not too complicated, database design and querying were easy. Finally, the computer seemed to have reached the stage of development of the modern car: the mechanic with his oilcan was no longer needed. Built-in 'wizards' compensated for lack of theoretical knowledge and querying a database could be as simple as searching for words in a text processor. One could even successfully complete certain tasks without knowing exactly what had happened.

But when exactly does a database become so complicated that standard facilities are no longer adequate? No red warning lights will flash! Software like Microsoft's Access had (and has) a tendency to mask real problems in database design, in particular in historical database design. Many historians discovered far too late that the design of their database did not meet their requirements

[25] Italics are all ours.

and / or the inherent structure of their source material, when the desired results failed to come out. Denley succinctly:

> "It has to be observed that the marriage of history and relational databases is one of convenience (some would say inconvenience) rather than design."[26]

The flexibility of the relational model in adding new data layers did not solve, of course, all the typical problems of historical computing. Introducing CLIO, Thaller had already pointed to the inherent fuzziness of historical data and the complex process of historical interpretation. If the entire process of historical data retrieval was left to the DBMS, somehow historical knowledge had to be incorporated. This was not easily done in the relational database environment itself.

The details of the relational model and related techniques of data modelling (like the Entity Relationship Model) were made widely known through the work of Lou Burnard and Charles Harvey, in several articles and in particular through Harvey's book on databases in historical research (Hartland and Harvey, 1989; Harvey, Green and Corfield, 1996; Harvey and Press, 1993). Burnard also recognised the complexity of representing historical reality in a computer. In designing a database, historians should start with a sound 'conceptual model'[27]), which comprised the real world objects, events and their relationships. Next, this model had to be mapped on to the sort of data structures a computer can deal with. He admitted that the integration of the different levels (from raw textual data to identified historical individuals and events) was not easy with standard software, but, in spite of that, he considered the relational model as the most viable solution. A lesser degree of refinement in automation might be acceptable: some information can be stored separately and administered by hand and the mapping from conceptual to physical model took place completely in the historian's head (Burnard, 1987, 1989, 1990).

3.4.2 Enhanced IT

3.4.2.1. Genuine historical software

From the beginning, the furthest-reaching position in the other line of thought has been taken by Manfred Thaller. At the first Westfield conference he presented a paper on methods and techniques of historical computation (Thaller,

---

[26] Denley (1994a), p. 35.

[27] A conceptual model is a map of the world whose data are administered in the database. It goes with an early stage in methodological system design. It exists usually outside the DBMS itself and is created with the aid of a CASE-tool or a diagramming package (e.g. nowadays Visio). The DBMS "does not know" about it. On basis of the conceptual model the database designer defines the physical model: the set of tables, with fields of a specific length and data type. Only the physical model is actually used by the DBMS in data management operations.

1987) which outlined his main ideas on this subject, though these were to be more fully formulated in later publications:[28]

1) Historians deal with problems not appearing in other disciplines, which should be controlled with a level of skill a historian can be expected to acquire without re-focusing his main research interest. So, the enhanced-IT view intends to make life easier for common historians by providing expert tools.
2) Historical data is to be administered as pieces of text, without any assumption about its meaning. Meaning depends on interpretation, which is a fruit of historical research. Therefore, data should be entered in a source-oriented way (keeping together in a single file what appears in a single source document), rather than in a program-oriented way. His definition of 'source-oriented' is, however, more inclusive than those in the previous section:

   > "Source-oriented data processing attempts to model the complete amount of information contained in an historical source on a computer: it tries to administer such sources for the widest variety of purposes feasible. While providing tools for different types of analysis, it does not force the historian at the time he or she creates a database, to decide already which methods shall be applied later." (Thaller, 1993b)

3) The typical historical database management system (like his CLIO / ??e??) would be a hybrid between a classic structured DBMS, a full-text retrieval system and a document retrieval system (which sounds more familiar in a time of XML-databases than twenty years ago), provided with some specific subject knowledge and inference mechanisms in order to enable historically meaningful data retrieval ('interpretation aware').
4) Such a system must be able to overcome differences in spelling (e.g. in surnames) and to handle data related to individuals in a careful way, allowing for research-based strategies for linking source fragments containing names to historical individuals (so-called 'nominal record linkage'). This would require the implementation of some knowledge in form of 'logical objects', containing rules for interpretation by the software.
5) Finally, it should take care of all other required transformations of data, for example, for the benefit of statistical analysis.

Thaller's view did not exclude the use of systems like dBASE or Paradox for uncomplicated data processing and simple data storage. However, he suggested that one should remain aware of the structural limitations of this kind software. His main concern was not about the software in itself. At the second Westfield conference he argued for a distinct *theory* of historical computing, a

[28] For a detailed, more technical description, refer to Thaller (1993a).

well-founded conceptual framework which would allow professional discussions about the peculiarities in historical data processing, firmly stating his belief in the fundamental difference between 'normal' and historical data processing (Thaller, 1989).

The obvious question to ask is, why ??e?? didn't sweep away its competitors in world of historical research like Microsoft Access did on the desktop. As Peter Denley noted in 1994 in his survey of the state of the art in historical database management, the power of the software has taken its toll. There is an almost infinite number of data structures possible; the tools to query, analyse, and manipulate the sources are powerful and sophisticated. User-friendliness was not made a priority, simply because in Thaller's opinion historical computing was a demanding science and that historians did themselves a great disservice if they made it look simpler. However, data preparation could be far less laborious with ??e?? than with a relational system, and not everybody needed to delve too deeply inside the tool set.

In addition, source-oriented data processing itself has attracted fundamental criticism. Many historians worry that purists who wish to represent the source electronically in a form that is as close to the original as possible, may be according a low priority to analysis, and may have a misplaced faith in the authority of text. Along with this line of reasoning the value of the source itself can be put into perspective as a mediated representation of the historical past (Denley, 1994a).

3.4.2.2. Historical tools on top of standard software

In addition to Thaller's strict source-oriented philosophy with far-reaching implications of non-standard software, several computing historians felt a need for an intermediate approach. They realised the limitations of commercial packages on the one hand and the value of general standards and proven methodologies in computer science on the other hand. This position inspired the creation of add-ons and reusable applications (or application frameworks) on top of standard software like relational database systems. These solutions were designed to comply with special requirements in historical data processing and, at the same time, would benefit from rapid developments around middle-of-road information technology.

It is hard to distinguish in this category between 'mere applications' and 'tools with a wider scope'. The delicate point is the way the added value (specific historical knowledge, algorithms or data structures) is documented and made explicit and available to a broader audience. Having a theoretical foundation together with some philosophy about how to serve historical research is an essential prerequisite. Lacking in this respect made these attempts fail in convincing fellow historians of its potentially greater value and classified them amid ordinary applications.

The list below is certainly not complete, but covers a few representative examples:

- One way of a realising this idea was creating a model application that clearly demonstrated how specific peculiarities of historical sources could be handled within, for example, a relational database environment, as Boonstra did for event-history (Boonstra, 1994a; Boonstra and Panhuysen, 1999; Boonstra, 1990). Gunnar Lind compared and analysed different structures for prosopography, suggesting a standard core database structure with the relational model in mind (Lind, 1994). Morris explored how standard applications could be combined efficiently for nineteenth century wills, hopping from one commercially available and well supported program to another and exploiting each application in areas of functionality in which it was strong and user-friendly (Morris, 1995).
- Welling studied intelligent data-entry strategies and user interfaces for highly structured sources and implemented his ideas in Clipper using dBASE-files for storage (Welling, 1993; Welling, 1992). Reflecting on his work he stood up for a distinction between 'history and computing' and 'historical computing'. The former concerned the contributions of computation to history. The latter has to deal with all the "grubby practicalities of hardware and software. It will have to deal more with applying what information science has taught us." Implicitly he criticised ??e??: "If we want to make software for historians, we must stop producing programs that require attending several summer schools before you can work with them" (Bos and Welling, 1995).
- Jan Oldervoll shared Welling's interest in historical tools with good interfaces, and created CensSys, a historical software package for analysing census data. Although CensSys was primarily designed as a fast and specialised tool for a specific kind of source, it was also based on clear ideas about interfacing between programs. Accepting its necessary limitations, it had provisions for delegating tasks to other programs; including an interface to ??e??(Oldervoll, 1992, 1994).
- Breure created SOCRATES, a framework on top of dBASE, consisting of a program library and tools (like program generators), that helped to build historical database applications with this popular database package. SOCRATES comprised not only software, but also a few guides ('grey publications') about historical data modelling. It particularly focused on problems with irregularities of source structures versus the strict demands of the relational model, and on the handling of text embedded factual data, like mentions of people, events and objects in wills and deeds (Breure, 1992, 1994a, 1994b).

3.5. Main topics in historical information science

Within the domain of historical information science, dozens of research themes have attracted attention from hundreds of historians and information scientists over the last 25 years. Some information problems were solved in the meantime; others have come up, while some other problems are still being discussed. There are also a few problems that have never attracted much attention, although they seem to fit very well into the domain.

At this point, we could present a detailed historiography of history and computing, as unfolded through numerous project papers in the proceedings of the several conferences, journals and workshop volumes. However, this has been done already by others: (Denley, 1994a; McCrank, 2002; Speck, 1994; Woollard, 1999). Nevertheless, it is useful to get an idea of all issues that have been discussed, rejected and (could have been) achieved within the domain of history and computing. The issues are grouped according to the kind of data to which they are related: textual data, quantitative data and visual data.

3.5.1. Databases and texts: From documents to knowledge

3.5.1.1. Databases and text in humanities research

Discussing databases and text in a single section may seem like a hasty attempt to cover quite dissimilar kinds of computer application in an overview, without doing justice to the complexities of each specific domain. The term 'databases' has strong connotations with relational database systems, tables and relatively small, highly structured chunks of data, while 'texts' is usually associated with full text and text analysis software for producing concordances, word counts, metrics for style, and, more recently, with text databases, supporting storage and publishing Web documents. 'Structuring' in classical databases means putting data in fields, while text is structured by 'marking up' information through the insertion of tags or other codes. For years these concepts have corresponded to different strains in the evolution of software, fostering an archipelago of database and text islands, where bridges were hard to build.

Recent technological developments, especially the advance of XML, have blurred this distinction. XML is widely used in business applications for the exchange of database data, which can be merged with running text, if required. XML-data may be stored in relational database management systems, or in dedicated object database systems. The landscape is changing now, and the islands are growing slowly together. Historical information, once liberated from its paper substratum, may subsequently "incarnate" in quite different shapes. So, the current state of technology is the first good reason for a combined approach.

Roughly speaking, ‘databases’ and ‘texts’ have also stood for different areas of humanities computing. We may safely say that database themes, in some form or another, dominate the majority of publications on historical computing, while text applications and text analysis abound in the field of literary computational studies. Of course, this image is stereotypical, but it still holds true and it would indeed justify two separate chapters. The second reason for combining is an attempt to gain a broad overall view of the entire field of storing and creating digital representations of historical and cultural documents. Facing the abundance of literature dealing with the peculiarities of database and text problems in the unique context of specific sources, it seems worthwhile take a few steps backwards, and, from a greater distance, to look for similarities and comparable methodological problems and strategies at a little higher level of abstraction. Because historical and literary studies have had their own publication channels, we should alternate between both fields if we want to discover parallel developments.

A few words of caution. In spite of this wide angle view, the reader should be warned that this section will not provide a balanced encyclopaedic description (for that purpose, one may, for example, refer to (McCrank, 2002)). Our main questions will be: What has been done so far? Where has it got stuck? What has to be done in the near future to ensure scientific progress? The underlying methodological question is, how historical data processing on this basic level of storage and transformation can be further streamlined, taking advantage of current developments in other disciplines, in particular information science and computer science. In search of answers, we will be selective, outlining primarily the developments at the historical side and looking for matching parallels in the literary field. On the other hand, some parts of this section may look too detailed and even trivial to humanities scholars. Because the text is not intended for this audience alone, a discussion of distinctive computational techniques will only make sense to people outside our field, if they are accompanied by a few introductory comments on characteristics of humanities studies in general.

3.5.1.2. Historical studies versus literary studies

Although both historical and literary studies use textual material as an object of study, methodologies and techniques are quite different in many respects, which has likewise resulted in differences in computational approaches. Literary research has been focused on critical editing and analysis of the texts themselves: the reconstruction of the lost common ancestor of surviving texts through the examination of manuscripts, the comparison of one text with another to discover textual variation, and the emendation of the text reconstructed, thereby removing errors. The analytical tradition has created a wide array of techniques and tools for textual analysis, varying from lemmatised word lists and concordances, to programs for stylistic and thematic analysis.

History is mainly based on written sources, which have, strictly speaking, also the form of texts. However, in contrast with literary studies, historical research problems pertain less to the texts themselves, but are more related to the historical reality beyond the documents handed down. Traditionally, historical questions will direct the attention to the factual elements in the texts, to person names, dates of birth and death, to belongings or to cargo carried in ships or to the departure and destination of voyages. Texts perceived in this way tend to lose their textual qualities and will be regarded as lists of data.

Where administrative documents are concerned, this point of view is mostly correct and efficient; however, the domain of historical sources is fuzzy and full of exceptions. Text features are indispensable links in the chain of historical interpretation and may be still be relevant in a later stage of research. They may reveal characteristics of unknown authors and of the history of the manuscripts themselves. A rare, but very good example is the use of cluster analysis, applied to strokes of letter forms occurring in undated manuscripts, written by the same scribe, in order to establish the probable date of completion (Friedman, 1992). Some highly relevant notes can be found in margins, a quite regular source may suddenly lose its structure at a certain point, or interesting information may be appended to regular data in an unexpected way, for example these cases in a British census list:

> "The ability to include apparently insignificant and microscopic detail from the census within the ??e?? DBMS has important macroscopic implications. For example, the refusal of a young woman to reply to the question on occupation in the 1881 census for Winchester, coupled with the comments of the enumerator, whose definition of her occupation as 'on the town' (implying prostitution) provides an important glimpse behind the curtain of the surviving sources – the enumerators' book, and towards an understanding of the process through which the original census schedules (which have not survived) were transformed into the documents we have today. Conversely, a two-line entry in the census which reads 'Assistant Classical Master/BA Trinity College Dublin' which is reduced by the editorial pen to 'Prof' helps the researcher to grasp some of the smoothing out process of categorisation which went to contribute to census statistics overall." (Burt and James, 1996)

However, traces of human life have not always been recorded in the form of lists. Charters, probate inventories, notarial deeds and wills form a category of sources which is not easily positioned on the sliding scale from structured data to running text. Both factual data and the surrounding text may be of interest and should be stored, therefore. In a subsequent stage, the former type of data will be used for quantitative analysis, while the latter kind of information may prove to be valuable for correct interpretation of individual cases. In 1995, *History and Computing* dedicated a special issue[29] to probate inventories and wills, which shows the struggle with these ambivalent data structure require-

[29] *History and Computing* 7:3 (1995), p. iv-xi, 126-155.

ments. Some researchers preferred to record the entire text structure, while others chose a middle path, entering data verbatim from the source, but without preserving the grammatical structure of the text. Software varied from special packages (table-based), to Paradox for Windows and at that time popular text database systems such as AskSam (Litzenberger, 1995; Morris, 1995; Overton, 1995; Schuurman and Pastoor, 1995; Webb and Hemingway, 1995).

In addition to documents that have been produced by administrative systems in the past, a substantial part of historical research is based on narrative sources, like chronicles, biographies, diaries, journey accounts, treatises, political pamphlets, and literary works. Here, historical, literary and linguistic scholars share a considerable amount of material, however, with distinctive intents, which has important consequences for information modelling and data processing.

The well-known diaries of Samuel Pepys (1633-1703)[30], a highly personal account of seventeenth-century life in London, are first of all of historical interest, but have also given rise to linguistic studies, and his work appears in literature courses, for example, in the broader context of studying the emerging modern expression of self-identity in the literature of his age. Historians will isolate and label historical events in the digital text, with mark-up for persons, places and dates, and prefer storage in a database for sorting and easy look-up, preferably with links to the original text. Alternatively, the voluminous text may be scanned first for themes of interest in order to locate relevant passages (Louwerse and Peer, 2002).Techniques like text mining, Topic Detection and Tracking, and Text Tiling, developed in information retrieval, could be helpful.

A study of Pepys' linguistic usage itself will require counting of words, phrases and linguistic constructions – Pepys sometimes used a kind of private code involving words from Spanish, French and Italian, obviously for reasons of concealment, hiding text from the casual browser. Stylometric analysis could help to cluster certain parts of these diaries in order to test hypotheses about the author's distinct characteristics in certain periods. A complex phenomenon as self-expression can be studied in a quantitative manner by applying content analysis techniques, searching for and counting key words and phrases related to this subject.

This example demonstrates that decisions about text encoding are far from easy, if a text is to be reused for various purposes. An important part of the discussion from both sides concerns what level of digital representation is adequate for historical and cultural source texts. In the historical field, this discussion has centred around the dichotomy 'source-oriented versus model-oriented'. In the domain of literary studies its counterpart is to be found in the debate on the nature of the critical text edition.

[30] See, for example, http://www.pepys.info.

3.5.1.3. The critical edition in the digital age

Background

Both history and literary studies share a reliance on high quality editions of textual sources. From the nineteenth century onwards history as a discipline has been based on an important tradition of printed critical source editions, including an extensive apparatus of footnotes, which explain and comment on the main text. The application of information technology has led to reflection on the nature of scholarly source editions. Some works have got a digital companion in addition to the printed book (e.g. as PDF-file or CD-ROM)[31]. In other cases, an on-line database will be an obvious solution, as with the material concerning the Dutch-Asiatic trade of the *Verenigde Oostindische Compagnie* (VOC)[32]. Voluminous printed editions, being the editor's lifework, are difficult to uphold. Moreover, information technology has liberated the critical source edition from the constraints of the printed book.

'Comprehensiveness' is an important goal, but cannot always be attained by time- and paper- consuming full-text editions. That is why the Institute of Netherlands History (*Instituut voor Nederlandse Geschiedenis*) decided to publish the correspondence of William of Orange, comprising approximately 11,000 letters, in the form of short summaries, carrying metadata and linked to digital images of the original sources. Another example is the edition of the Resolutions of the Dutch States General. An electronic edition with full-text search facilities would have been attractive (searchable on-line as with the *Papers of George Washington*, president from 1789 to 1797[33], and with the scientific archive of Samuel Hartlib, c.1600-1662[34], which has been published as a CD-ROM edition of text images with transcriptions), but bare text-retrieval software doesn't handle these kind of seventeenth century texts very well, because they contain many spelling variants and terminology quite different from the concepts modern historian are looking for. Therefore, for the time being, one of these projects has settled for a compilation of summaries in modern Dutch (Haks, 1999). More systematically, a useful classification is that into 'digital facsimiles', 'digital editions', and 'digital archives' (Thaller, 1996).

Digital facsimiles

A digital facsimile provides access to an individual hand-written or printed source text, by means of scanned images, a complete transcription, and a linked database of persons, locations and concepts (specifically 'time') mentioned in the

---

[31] For example, *Kroniek van Peter van Os. Geschiedenis van 's-Hertogenbosch en Brabant van Adam tot 1523*, A.M. van Lith-Droogleever Fortuijn, J.G.M. Sanders & G.A.M. van Synghel ed. [Instituut voor Nederlandse Geschiedenis], Den Haag (1997).

[32] Instituut voor Nederlandse Geschiedenis: http://www.inghist.nl/Onderzoek/Projecten/DAS/EnglishIntro.

[33] American Memory Project: http://memory.loc.gov/ammem/mgwquery.html.

[34] The Hartlib Project: http://www.shef.ac.uk/~hpp/index.html.

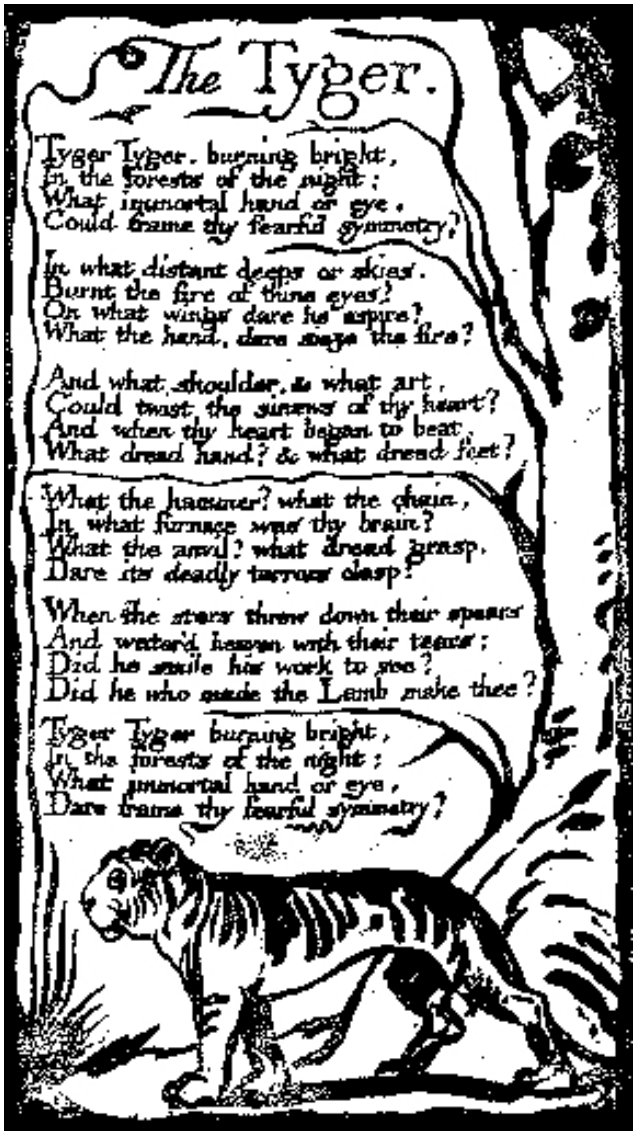


Figure 3.1. The Blake Project: 'The Tyger', *Songs of Innocence and of Experience* copy C (Viscomi).

text. Optional extras for historical sources are other tools, which document former calendar systems, currencies, and specific terminology. Within literary studies, the term 'image-based computing' has a special meaning. It may be said to descend from the so-called social or materialist theories of textual production advanced by such scholars as McKenzie and McGann in the early eighties, coupled with the means to create (relatively) high-quality and (relatively) low-cost digital facsimiles of documents. A wide array of special applications is grouped around this concept, which Kirschenbaum in his introduction of a special issue of *Computers and the Humanities* has called 'venue for representation' (Kirschenbaum, 2002).

A nice example is the work of William Blake, an eighteenth century poet, who illustrated his own work in watercolours, later printed from a copper plate. In a certain sense, Blake created an eighteenth-century multimedia presentation: he used printing as a mode of production rather than reproduction, etching himself and combining text and illustrations on the plate for the first time rather than reproducing a pre-existent page design. Editorial problems arise with numerous, later impressions, often of an inferior quality, all more or less different. It may be obvious, that a mere transcription does not do any justice to the work's artistic qualities and that the great number of variants are not easily reduced to a single edition. A natural solution was a digital facsimile edition, with trustworthy reproductions of the illustrated text and full transcription in SGML. An advanced image manipulation tool enabled resizing to actual size at any monitor resolution. The images could be examined like ordinary colour reproductions, but could also be displayed alongside the texts, enlarged, computer enhanced, juxtaposed in numerous combinations, and otherwise manipulated to investigate features (such as the etched basis of the designs and texts) that were previously imperceptible without close examination of the original works (Viscomi, 2002).

Digital editions

A digital edition goes a step further, providing access to different versions of the text, realising the aims of the critical edition in digital form. A good example, from the literary domain, is *The Canterbury Tales Project* (De Montfort University) which aims to:

- Publish transcriptions of all the manuscripts and early printed books of Chaucer's Canterbury Tales into computer-readable form (eighty-four manuscripts and four printed editions survive from before 1500!).
- Compare all the manuscripts, creating a record of their agreements and disagreements with a computer collation program.
- Use computer-based methods to help reconstruct the history of the text from this record of agreements and disagreements.

- Publish all the materials, the results of the analysis, and the tools which were used by the researchers (materials are available both on-line and on CD-ROM)[35].

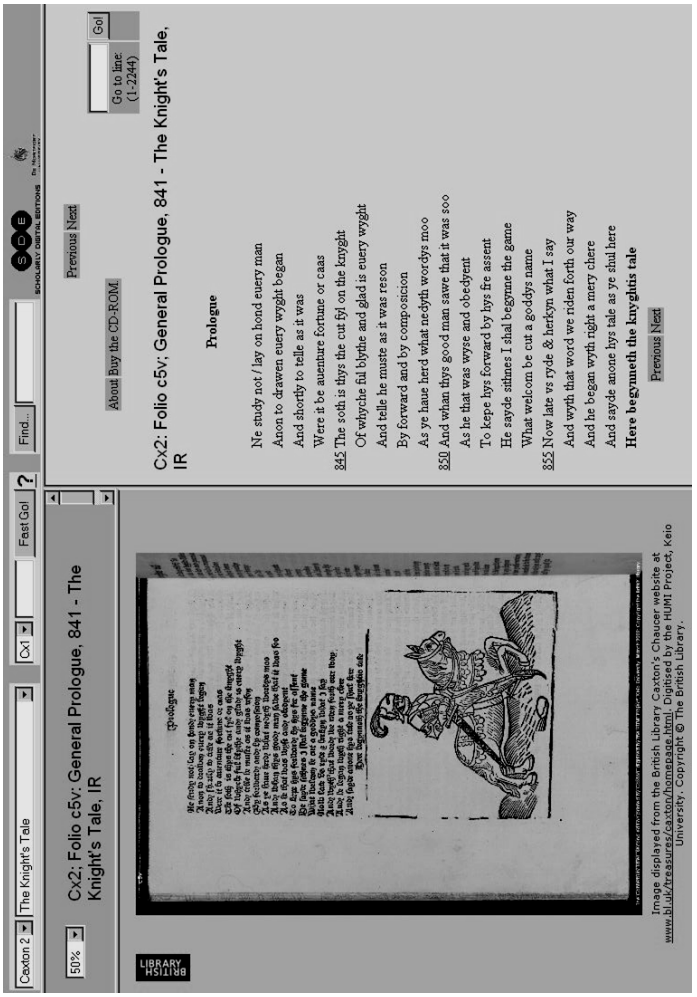


Figure 3.2. The Canterbury Tales Project: Facsimile with transcription of the second Caxton edition. The search box with Find-button will produce a KWIC-index for a given key word.

[35] The Canterbury Tales Project: http://www.cta.dmu.ac.uk/projects/ctp/index.html.

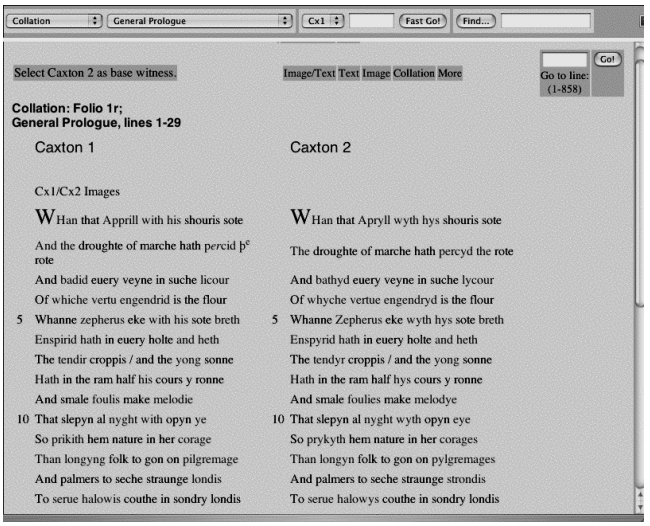


Figure 3.3. The Canterbury Tales Project: Word by word collation of whole text of both Caxton editions, with all differences highlighted in red.

Digital archives and digital libraries

A digital archive (or virtual archive) is characterised by making a large collection digitally available through a variety of retrieval mechanisms, linking different databases behind a uniform user interface, with additional tools for analysing information (e.g., mapping software) and offering some personalization. Because of the larger scale the granularity of disclosure will vary greatly. The term has strong connotations with administrative documents, which shows to full advantage in projects like the computerisation of the *Archivo General de Indias* (González, 1998)[36], which holds all documents concerning the Spanish administration in the Americas and the Philippines, the *Duderstadt Project* (Aumann, Ebeling, Fricke et al., 1999), which is developing a computerised version of the files of the municipal archive[37] and the digitalisation of the medieval archive of Regensburg (the *Fontes Civitatis Ratisponensis* – FCR)[38].

However, the term is also used for different kinds of collections, e.g., the Thomas Jefferson Digital Archive (containing mainly letters), the Codices Electronici Ecclesiae Coloniensis (CEEC – digitized manuscripts of the church

[36] Archivo General de Indias: http://www.clir.org/pubs/reports/gonzalez/contents.html.
[37] Duderstadt Project: http://www.archive.geschichte.mpg.de/duderstadt/dud.htm .
[38] Fontes Civitatis Ratisponensis: www.fcr-online.com .

of Cologne) and Prometheus (a digital image archive), and the World of Dante, a hypermedia visualization of the work of the famous poet (Parker, 2001)[39].

There is only a vague borderline between digital archives and digital libraries such as *Perseus*. Organizations like the *Council on Library and Information Resources* (CLIR)[40] and *Digital Library Federation* (DLF)[41] do care for both, subsuming them under the comprehensive notion of networked digital repositories.

Electronic textuality and the New Philology

In the literary domain, the advances of information technology have given rise to the concept 'electronic textuality', grouping together issues about digitized material and the potential of hypertext techniques for electronic text editions (Abram, 2002). A great deal of current electronic texts has been encoded according the guidelines of the Text Encoding Initiative (TEI). The TEI guidelines uses descriptive mark-up, e.g. <title>, <chapter>, <paragraph>, in contrast with mark-up that specifies formatting procedures or presentation through fonts and colours. The TEI convention has been implicitly based on an underlying text model, which defines a text as "a hierarchy of content objects": as composed of nesting objects such as chapters, sections, paragraphs, extracts, lists, and so on. Although practical and appropriate in many cases, it has provoked extensive discussions, in particular with regard to overlapping hierarchies (depending on multiple points of view) and texts that lacked a clear hierarchical structure, such as poems (Ide, 1995; Renear, Mylonas and Durand, 1993).

By the mid-1980s, however, a more fundamental criticism came from textual scholars like McGann and Shillingsburg, who viewed a text primarily as a product of social interaction between a number of agents: author, editor, publisher, composer, scribe and translator (Schreibman, 2002). This has started a debate about the form and function of the critical text edition. The so-called New Philology has questioned the role of the editor in favour of the position of the reader. It fits into the post-modern thinking against all forms of authority and pays more attention to the historical situations of texts, their function in time and place, and to the interaction with their social context (the "textual turn"). It no longer sees different versions of a text as witnesses of a lost original, which has to be reconstructed from variants, found in extant copies. Not a reconstructed text, but a diplomatic transcription of texts handed down, has to be the basis of an edition (Kobialka, 2002; Ott, 2002; Robinson, Gabler and Walter, 2000).

---

[39] Thomas Jefferson Digital Archive: http://etext.lib.virginia.edu/jefferson/. CEEC: http://www.ceec.uni-koeln.de/. Prometheus: http://www.prometheus-bildarchiv.de/. The World of Dante: http://www.iath.virginia.edu/dante/.

[40] CLIR: http://www.clir.org/.

[41] DLF: http://www.diglib.org/.

Of the scholars working in this area McGann and Landow have been especially influential. McGann's ideas about 'social editing' and 'hyperediting' emphasise the value of hypertext and hypermedia in relation to the social aspects of literary texts (McGann, 1991, 1992, 1995, 2002). Landow argued that "the dispersed text of hypertext has much in common with the way contemporary, individual readers of, say, Chaucer or Dante, read texts that differed from one another in various ways" (Landow, 1996?). Editing becomes an ongoing process by means of collaboratories, where readers can play an important role in on-line annotation.

The consequences of this new paradigm for editorial practices and tools have been clearly expressed in the Electronic Variorum Edition of Don Quixote, which has as primarily goal "to develop a replicable program that permits the creation of online critical editions as hypertextual archives, using the *Quixote* as test bed." Strictly speaking, it is no longer an edition, but a dynamic, hypertextual archive composed of a series of databases with special tools, such as a text collator, a viewer for displaying digital images of the text and transcriptions side by side, annotation and update facilities for texts and stored information objects (Urbina, Furuta, Goenka et al., 2002).

Of course, the liberal publishing of documents and empowering the reader will create new problems: digitisation should not become a substitute for scholarship, and the new means of cheap distribution poses the question how to select documents and where to stop. Moreover, as Prescott has remarked in the context of the *Electronic Beowulf*, the impression has always been that digital images will be free or at least very cheap, thanks to governmental grants. The free ride will come to an end, when digitisation projects have to recover their costs (Prescott, 1997). This uneasiness has led to new methodological solutions, as formulated, for example, by Vanhoutte (the *Streuvels Project*), who made a distinction between the archival function and the museum function.

The archival function is responsible for the preservation of the literary artefact in its historical form and for documenting the historical-critical research of a literary work. The museum function pertains to the presentation in a documentary and biographical context, intended for a specific public and published in a specific form (Vanhoutte, 1999).

It is, of course, beyond the scope of this report, to draw up a balance. As for text projects it appears that hypertextual techniques have captured a core position in the editorial field. New forms of source editing have been established, varying from digital facsimiles to digital archives and digital libraries, all equipped with an array of dedicated tools, changing the traditional role of the editor, and empowering the reader, but without making the editorial rigor obsolete.

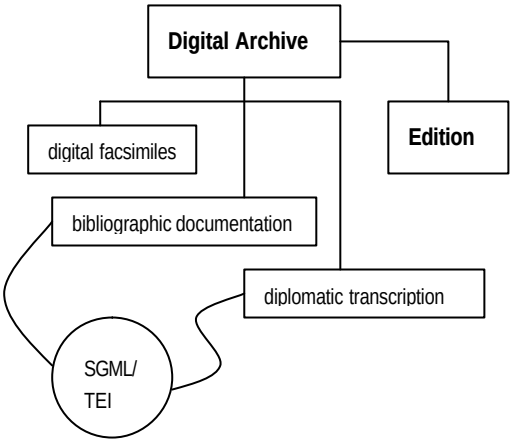


Figure 3.4. The Archive/Museum model by Vanhoutte.

For historians, the critical edition is the most complete way of making material available, in comparison with other forms of digitisation discussed below, and, therefore, it has been presented first. Application of computer technology creates new capabilities in disclosing information, however, automation should be used also to make the process of disclosure less time-consuming (e.g., the previously mentioned letters of William of Orange and the Resolutions of the Dutch States General). Therefore, progress with this kind of publication is closely linked with solving problems of digitising larger quantities of historical data for analysis purposes.

3.5.1.4. Digitising historical data for analysis

Unfortunately, the majority of historical research material will never be available in the form of a critical edition. Although libraries and archives on the one hand and researchers on the other hand may co-operate more in digitising, a considerable effort in this respect will be left to historians themselves. In this category project aims and constraints are usually quite different from those with critical editions discussed above. Nobody wants to put more effort in data entry and data processing than necessary, however, the complexity of historical research, which sometimes comes close to the work of a detective, may make it difficult to determine what 'necessary' exactly means. Thaller has summarised neatly the nature of problems with representing historical sources in a discussion on digital manuscripts:

> "Speaking on the most general level, we consider a text to be "historical", when it describes a situation, where we neither know for sure, what the situation has been "in reality", nor according to which rules it has been converted

into a written report about reality. On an intuitive level this is exemplified by cases, where two people with the same graphic representation of their names are mentioned in a set of documents, which possibly could be two cases of the same "real" individual being caught acting, which, however could also be homographic symbols for two completely different biological entities.

At a more sublime level, a change in the colour of the ink a given person uses in an official correspondence of the nineteenth century could be an indication of the original supply of ink having dried up; or of a considerable rise of the author within the bureaucratic ranks.

Let us just emphasize for non-historians, that the second example is all but artificial: indeed the different colours of comments to drafts for diplomatic documents are in the nineteenth century quite often the only identifying mark of which diplomatic agent added which opinion." (Thaller, 1996)

A clear formulation of the research problem and limitations in time and money will usually dictate practical solutions with regard to entering 'what' and 'how', frequently overruling more theoretical considerations of potential reuse. Although these problems with data representation are generally well known in the historical community, we shall briefly review the implications surrounding these problems:

– *The orthography of data*

The spelling of names of persons and places, the denomination of currencies and the amounts expressed in old weights and measures may require further standardisation in order to be analysed statistically. An important decision is at which stage this standardisation has to be carried out and how the coupling with the original information is preserved. The main problem behind it is the linking of source data to the historical entities as we know them: for example, to which geographic location does a certain placename refer, and which variant names correspond with the same historical person? Precisely because such answers may require further research, name standardisation should be avoided at data entry. Preferably, it should be possible to repeat the standardisation process automatically with different rules afterwards. A closely related problem is the reading of historical data (i.e., the particular way of understanding what is written). Sometimes individual data are difficult to separate from their context. A qualification as 'smith' may indicate an occupation or a surname: what should be stored where?

– *The data model*

A major point of discussion has been, whether the design of historical databases should reflect the model of the historical situation as envisaged by the researcher, or whether it should reflect the structure of the sources themselves. A historical source is a complex network of information, usually well understood by the historian, but difficult to represent in conventional storage systems. Although relational database systems allow the implementation of hierarchical data models (typical for texts), several kinds of historical material do not

fit into this scheme easily. In addition, data structures may vary over a group of sources of the same kind, or even worse, within a single source. This makes conventional data modelling[42] a precarious matter in those cases (for examples: (Burt and James, 1996)).

However, it becomes much less of a problem, when the research model, rather than the structure of the source, is mapped onto the database design, e.g., an economic market with suppliers, goods and consumers, or employment in an occupational group (Bradley, 1994). The particular form of choice made depends largely on the nature of the project and the refinement in computer assistance. The tension between the complexity and the irregularity of the data structure of historical sources on the one hand, and the rigid nature of the relational database model in combination with analysis purposes on the other has fuelled the discussion about source-oriented software in the 1990s decade (see below).

– *Linking data with the original source text*

In historical discourse historians are used to refer painstakingly to the information's place of origin in the underlying sources. This approach has also affected the architecture of historical computer applications that are used for research purposes. Although source data have to be converted to a form suitable for specific processing (e.g., statistical analysis – see below) and will appear at different levels of aggregation, many projects have demonstrated the need to preserve links to the deepest level, i.e., to the original source text. This can be realised by simply linking scans of the source document to the respective records in the database, or by adding the source text in transcription. The latter may be realised through the extensive use of text fields, or with the help of text look-up mechanisms and other built in applications (Breure, 1995a, 1995b).

Source-oriented versus model-oriented

As described in Section 3.4, in the 1990s a part of the historical community had become aware of the typical problems inherent to historical data processing. Harvey, Press, Greenstein and Burnard have done a lot to make historians acquainted with the theoretical principles underlying relational database systems (RDBMS) (Burnard, 1989, 1990; Greenstein, 1989; Harvey and Press, 1992, 1993, 1996). The discussion about "whether or not to use an RDBMS" tended towards a distinction between different kinds of projects. Some rules of thumb were formulated to help fellow historians make the right choice.

Denley did so in his balanced article on historical database design in the 1990s (Denley, 1994a). The model-oriented approach is for researchers with specific questions, using regular sources, accepting some arbitrary decisions about data and with quantitative analysis at the forefront of his/her intentions. A tight schedule and mainstream tools may be other arguments to choose this

[42] Data modelling: designing the structure of the database.

path. On the contrary, the source-oriented approach is more appropriate when the historian places high priority on maintaining the integrity of the source, wants to treat his material both as text and structured data, aims at a database that comes close to a source edition and has more time to spend on complex tools.

About the same time this point of view was demonstrated in practice by Bradley in a reconstruction of the British medical profession. He chose deliberately a relational database, accepting that the model would be a simplification of the historical reality, and nothing more than an attempt to replicate the structure of employment (Bradley, 1994). An RDBMS is fine, if a project aims at data analysis, rather than at source analysis, the data are open to aggregation and therefore to be used in statistical analysis and relationships between the database objects can be described as 'one-to-many'.

Two years later Burt and James praised the superiority of ??e??, its fluidity and flexibility at data entry and emphasised the macroscopic relevance of microscopic details:

> "Thus it is argued here that source-oriented models set a benchmark for historical studies in computing. This benchmark is above the level attainable by the rigid and exclusive technique of the relational database. There may indeed be a different mindset in operation in the source-led and source-oriented approach of historians when compared to the mindset of certain practitioners of computing and data mining in business and computer science. The microscopic detail in the historical source can prove to be of key significance in macroscopic results." (Burt and James, 1996)

For the remainder of this period both streams have run in parallel. It is beyond the scope of this report to quantify the market share of each.[43] Choice for the one or the other has largely depended on the particular interests and preferences of the researchers as described above.

The situation in about 2000 was summarised well by a Swiss PhD-student, Christian Folini. Having a good overview of current computer practice in the historical field, he discussed the needs of historical (graduate) students on basis of a small e-mail survey (Folini, 2000). He found a preference for relational databases, particularly for Microsoft Access. His rules of the thumb for selecting a strategy form a correlate to the scheme as described by Denley. He also found a group of researchers who used relational database and worked simultaneously with full-text systems. His own research was about female mystics in southern German convent of the thirteenth and fourteenth centuries, encompassing also full-text material. Finally, he based his solution on Access, with text excerpts entered in text fields of the relational database. His complaints

[43] This does not mean, however, that relational desktop databases as FoxPro, FilemakerPro and Microsoft Access have been opposed to ??e?? alone. A minor position in the source-oriented camp has been hold by full text systems such as TACT, Idealist, Wordcruncher, and Atlas.ti.

were about difficulties in estimating the time required for applying information technology, the lack of technical support, the rapid succession of software releases, and the technical limits of the basic versions of software, usually installed at universities, and concluded with hope on the unifying role of XML. His account is interesting, because his approach was open-minded and it demonstrates a serious lack of reliable and dedicated tools, especially for a younger group of researchers who are willing to apply computer techniques, but who have to cope with severe constraints in time and lack of straightforward methodology.

Multiple options: What is the problem?

How should we assess this situation? Can we conclude that the historical community has found appropriate practical solutions for natural difficulties inherent in the historical craft, and that there is not any serious problem at all? The answer is both Yes and no. 'Yes', because, in daily practice – as Folini himself demonstrated – one can manage to get on with the various options regarding source-oriented and model-oriented. 'No', from a methodological point of view, because the gap between model-oriented and source-oriented requires far-reaching decisions in an early stage of a project. A versatile approach would be more attractive, leaving open various options of database processing, statistical analysis and text analysis, without enforcing redundancy in the machine-readable source material, which easily happens when the same material must be encoded in different ways for different packages. In addition, our research agenda and related assumptions with regard to encoding and storage should be reversible within reasonable limits. This will be feasible only if the effort in digitising and data entry can be considerably reduced by applying in this stage computer techniques on a larger scale. Processing data requires structure. This may be created manually, but it is a challenge to discover to what extent information technology can be deployed for that purpose. There is an additional reason why 'the gap' is to be considered harmful: accepting the current situation without any further action will also widen an other gap, the one between historical information science and computer science, thus leaving ample ground for many projects to keep reinventing the wheel.

These ideas are far from new and return us to the very beginning of the discipline. In 1980 Thaller outlined already the ideal trajectory when he introduced CLIO (the original name of ??e??): historical data processing should move through different stages, starting with a data representation as close as possible to the original source, to more practical formats suitable for analysis and retrieval (Thaller, 1980). ??e?? comes with a so-called 'logical environment', comprising several rule-based algorithms for pre-processing raw source material.[44]

[44] See for example the Introduction of the on-line tutorial: http://wwwuser.gwdg.de/~mthalle2/manual/tutorial/intro.htm

The fact that after almost twenty years since its introduction not every historian is working with ??e?? cannot be explained by theoretical inadequacies with regard to the practice of historical research. Neither has the system been criticised for its lack of power. However, user unfriendliness, a steep learning curve, the feeling of a black box deploying a technology far away from generally accepted standards and mainstream computing have kept many potential users away, especially in a community that has not been particularly fond of computers at all (Denley, 1994a; Everett, 1995). It must be admitted that this criticism is not (fully) justified. Much has been based on blunt misunderstanding and lack of interest. In the meantime, ??e?? has evolved as well: it is now web-enabled and has learned to speak XML. ??e?? uses a very general data model, related to the semantic network, which can represent XML data structures as a subset (a capacity which is not immediately clear from the outdated documentation available on the website).

Irrespective of whether one wants to use ??e?? or not, the problem doesn't seem to be that we have no idea how to bridge gap. Greenstein and Burnard pleaded for joining the parallel trajectories of what they called the 'textual trinity' of (1) printing, publishing, and word-processing; (2) linguistic analysis and engineering; (3) data storage, retrieval, and analysis. They showed how TEI solutions could be used to create machine-readable transcriptions comprehensible for different applications. They demonstrated the power of TEI in encoding multiple and even competing interpretations of text (Greenstein and Burnard, 1995). The main drawback of this solution is the effort required: in most cases, there are simply more sources to be used than we are able to encode (remember: we are not envisaging here critical editions, but datasets for a particular research project).

Looking backward and tying up loose ends following from the variety of arguments above, the core question seems to be: how can we create structure in historical material in a way that is:

1) appropriate to the source's *complexity*,
2) *modular*, i.e., split into discrete and transparent steps, each preferably well-documented and clearly modelled, which
3) adheres to *standards* and
4) allows a *critical choice* of the best tools / techniques available (where 'best' implies, among others things, 'appropriate', 'well documented' and 'user-friendly'),
5) without spending an unwarranted amount of *time* either to manual encoding or to developing complex technological solutions.

Commercial database systems have satisfied criteria 2-4, but cannot easily represent the complexity of structure (criterion 1), and will therefore only be a good solution in those cases where structure is relatively simple and regular or when it has been defined as such within the project's framework. Scanning and manual XML encoding of full-text is unrealistic for many mainstream histori-

cal projects, due to point 5. This last requirement has been precisely the major motivator for developing ??e??, although this system seems to have suffered from an image of being bound to a non-standard, monolithic solution, thus failing on criteria (2), 3-4.

3.5.1.5. Automatic structuring

Creating structure and linking the structured text to semantic models (e.g., authority lists of persons, places etc.) are essential for historical data processing, however, it is desirable to automate this process to a large extent. Recently, technological solutions for transforming raw text into encoded information by means of (semi-)automatic techniques have made good progress. A few long-lasting, large-scale projects seem to be well on the way to satisfy all of the criteria mentioned above to great extent. They draw upon a combination of techniques from different domains, like natural language processing, text mining, speech recognition and corpus linguistics. These solutions are mostly not (yet) disseminated in a form of ready-made tools that can be easily used by others, like concordance programs and statistical packages. The variety of original methodological contexts makes it far from easy to decide 'what' to use 'when' and 'how' in specific historical research. Although they certainly fall outside this category of 'digitising for practical purposes', they indicate a promising direction for future methodological and interdisciplinary methodological research.

Perseus

The *Perseus Digital Library*[45] is not only interesting as an on-line source of a wealth of information, but also because of methodological aspects, in particular the transfer of methods and techniques from one domain (antiquity) to another (modern time, the history of the city of London), covering different languages (Greek, Latin, English, Italian, Arabic), and its intention to formulate a more widely applicable strategy of digital collection building, together with a generalised toolset. *Perseus* has a philosophy of starting with simple mark-up (concerning morphology, or the tagging of proper names of places and persons on the basis of different authority lists), successively taking advantage from each information layer, without immediately striving for a perfectly encoded text. Crane has well documented this strategy (Crane, Smith and Wulfman, 2001):

> "Automatic tagging takes place in two steps, of which only the first has been fully implemented. In the first step, we look for proper names but make no attempt to resolve ambiguities. We tag "Oliver Cromwell" as a personal name but do not try to determine which Oliver Cromwell is meant, nor do we look for instances such as "the Oliver Cromwell" (which might refer to a building or institution named after the historical figure).

[45] Perseus: http://www.perseus.tufts.edu/.

> Once possible proper names have been tagged, there are various strategies to analyse the context and rank the different possible disambiguations. Our energy at this stage has focused on acquiring and, where necessary, structuring the data that we have had to enter ourselves.… The human editor could also enter at this stage, going through the automatically tagged text. Ideally, the editor would find most features properly identified and would have to intervene in only a small percentage of cases.
>
> But even without disambiguation or hand-editing, we have been surprised at how useful the subsequent electronic environment has proven. We consider this to be an important finding in itself because the performance of a system without clever disambiguation schemes or expensive hand editing provides the baseline against which subsequent improvements can be measured. Our experiences suggest that both clever disambiguation and hand editing will add substantial value to documents in a digital library, but, even failing those two functions, the automatically-generated tags can be employed by useful visualization and knowledge discovery tools." (Crane, 2000)

From the beginning *Perseus* has adhered to standards (SGML, XML, TEI, relational databases). Originally it did not spend much effort in programming (one programmer until 1994) and developed only one important piece of software: a rule-based system to analyse the morphology of inflected Greek words (later extended to Latin and Italian). Gradually, more tools have been created and made publicly available on-line.[46] By around 2000 the toolset comprised sophisticated full-text searching facilities, the creation of links among documents in the system, extraction of toponyms and the automatic generation of maps, discovery of dates and the dynamic display of timelines, the automatic implicit searching and discovery of word co-occurrence patterns, and linkages to morphological analysis (Rydberg-Cox, Chavez, Smith et al., 2002; Smith, Rydberg-Cox and Crane, 2000).

These tools enable the generation of a knowledge layer, consisting of repertoria, large collections of metadata and comprehensive display schemes. An interesting example in this context is the inference step from text to knowledge through collocation analysis. For this purpose the technique of Topic Detection and Tracking (TDT), developed in information science under guidance of DARPA, was tested and adapted to historical needs.[47]

TDT aims at developing techniques for discovering and threading together topically related material from streams of data such as newswire and broadcast news. TDT systems will aggregate stories over a span of several days into single event topics. The most significant problem in adapting TDT methods to historical texts is the difficulty of handling long-running topics. Many historical documents discuss long-running events, and many users will wish to browse digital libraries at a scale larger than events of a few days' length.

[46] Tools available through the tool page of Perseus itself, and through the Stoa consortium: http://www.stoa.org/).
[47] Topic Detection and Tracking: http://www.nist.gov/speech/tests/tdt/.

Moreover, historical texts tend to be discursive, not broken into discrete date units, and digressive. Even if there is a main linear narrative, a historian will often digress about events from before or after the main period, or taking place in another region. These digressions, of course, may themselves provide information about other events. Last but not least, date extraction is far from easy, amongst others because dating schemes other than the modern, Western Gregorian calendar. As a solution place-date contingencies were calculated and several measures of statistical association were tested, to find the best ranking of events in the presentation of query results (Smith, 2002).

The Integrated Dutch Language Database (INL)

The *Instituut voor Nederlands Lexicologie* (INL – Institute for Dutch Lexicology) in Leiden[48] administers the Dutch linguistic heritage by automatically encoding text and storing millions of words in databases: the Integrated Language Database covering the eighth to the twenty-first centuries, and a subcorpus of the international PAROLE database (a mixed corpus of newspapers text, newsreel footage and periodicals). It goes without saying that processing texts on such a large scale requires extensive automation. Raw texts are tagged as much as possible automatically using TEI, using PoS-taggers (PoS: Part-of-Speech) for word classification[49] and rule based programs for sentence splitting. This technology has been developed in natural language processing and comprises different strategies (e.g., rule-based tagging, memory-based learning and Markov-models) (Depuydt and Dutilh-Ruitenberg, 2002; Does and Voort van der Kleij, 2002; Dutilh and Kruyt, 2002; Raaijmakers, 1999).

Also of interest in this area is the testing of this technology on a wide variety of historical material. General algorithms are tested for appropriateness for domain specific purposes, thus leading to more refined solutions. This is mostly related to characteristics of the Dutch language, and has resulted, for example, in the development of the PoS-tagger and testing of different underlying models. We see the same strategy here as in the *Perseus* project: more general technologies are systematically tested and transformed into suitable tools.

Other examples

There are several converging, more detailed research lines in creating structure automatically. As mentioned before, from the beginning ??e?? has had a 'logical environment' with rules and procedures to convert historical data sets in this way. Currently, it is used in several large-scale digital archive projects like CEEC, Prometheus, Duderstadt and Regensburg (see above). The application of automatic procedures is reported to transform the raw source text into an encoded data set by means of rule-based editing and semantic parsing (Kropaç, 1997).

[48] Instituut voor Nederlandse Lexicologie: http://www.inl.nl/.
[49] For an overview, refer to http://www-nlp.stanford.edu/links/statnlp.html.

Text-image coupling is essential for facsimile edition. The linking should be precise. Lecolinet et al. reported progress in automatic line segmentation of scanned hand-written manuscripts. They developed a semi-automatic approach that lets the user validate or correct interactively transcription hypotheses that are obtained from automatic document analysis. This capability facilitates interactive coupling by pre-segmenting manuscript images into potential line (or word) areas. As hand-written documents often have quite a complex structure, it is generally not possible to process them in a fully automatic way. Consequently, user validation and correction is needed with most documents (and especially with modern hand-written manuscripts) (Lecolinet, Robert and Role, 2002).

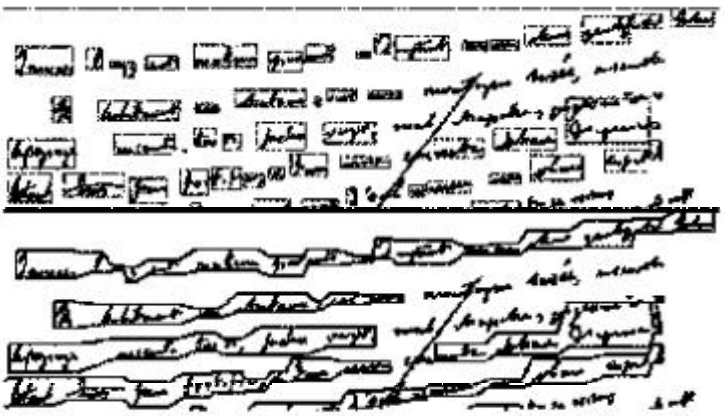


Figure 3.5. Automatic line segmentation in scanned manuscripts (Lecolinet, Robert and Role, 2002)

From 1997 to 2001 the Institute of Netherlands History (ING) and the NIWI have co-operated in retro-digitising the printed volumes of the *Repertorium van boeken en tijdschriftartikelen betreffende de geschiedenis van Nederland* (Repertorium of publications on Dutch history), covering the years 1940-1991. Being scanned, the bibliographic information had to be corrected, completed and split into database fields (the repertorium is now a database in Pica-format, which is used by a substantial number of ministry libraries, mainly in the Netherlands and Germany). In particular the older volumes caused all kinds of problems due to irregularities in typography as well as typical 'book conventions', such as cross-references to other sections, well understood by readers but very impractical for database storage. The large quantity of pages to be processed made the use of automated procedures paramount. The entire project has comprised several experiments in (semi-)automatic structuring, among other things,

by implementing sub-processes in Perl and using regular expressions in more advanced text editors as TextPad.[50]

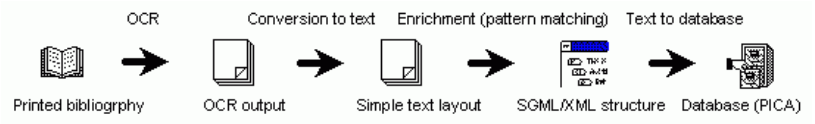


Figure 3.6. Repertorium Project (NIWI): Conversion path.

3.5.1.6. Nominal Record Linkage

Any discussion on texts and databases is incomplete without a discussion of a particular technique of distilling knowledge from data: nominal record linkage, which links occurrences of names in different source texts to historical individuals.

The obvious first step in constructing historical knowledge is identifying the individuals, who left their traces in historical sources. The extensive literature about this subject, dating back to the early 1970s (Winchester, 1970; Wrigley, 1973), deals with questions as how spelling variants in sources are to be standardised, to what degree linkage can be automated, and what level of relative confidence is acceptable with automated procedures. A final consensus has not yet been reached, in spite of thorough debates, experiments and testing. For a great deal this may be explained by the variety of sources and historical problems involved, covering different times and cultures, thus creating an endless range of peculiarities: poll books, census and tax registers, baptism and death records, used for a diversity of research purposes as studying political behaviour, land holding, family reconstruction, life course analysis, regional demography, and prosopography (i.e. using individual data to describe the history of a group[51]) (Adman, 1997; Adman, Baskerville and Beedham, 1992; Davies, 1992; Harvey and Green, 1994; Harvey, Green and Corfield, 1996; King, 1992, 1994; Ruusalepp, 2000; Tilley and French, 1997; Vetter, Gonzalez and Gutmann, 1992; Winchester, 1970).

Harvey and Green, studying the political behaviour of eighteenth century inhabitants of the City of Westminster on the basis of poll books, have given a few examples of the problems one may come across:

> "Some voters moved house during the period; others changed their jobs; surname spelling was inconsistent; data drawn from printed or manuscript copies of the original poll books contain transcription errors; and some data have been lost. Each of these may lead to a failure to link records which relate to a person. Moreover, the practice of naming sons after fathers, and of those sons

[50] Information from D. Stiebral. For project information: http://wwwoud.niwi.knaw.nl//nl/dd_nhda/projects/proj_rep.htm.

[51] For the different shades of meaning of prosopography, refer to the introduction to the special issue about this theme in *History and Computing* 12:1(2000).

> inheriting their fathers' estates, gives rise to the possibility of establishing false links between distinct voters. This possibility may also arise by personation, as well as by the coincidence of two distinct voters in successive elections sharing common names, addresses, and occupations." (Harvey and Green, 1994)

Although nominal record linkage is mainly discussed with regard to administrative sources, the problem is not restricted to this category alone, and has its counterpart in literary texts as well. In a prosopographical study about the struggles between Norwegian factions in the period of civil wars (1130-1240) Opheim has constructed a database from the kings' sagas, encoded in XML. Recording the participants in the civil wars requires the identification of personal names, place names and group names. A person may be referred to by his given name and a patronymic (e.g. Dag Eilivsson), or with some signifiers of status or office (king, archbishop), or individual characteristics (Guthormr Grey-Beard), but any combination may occur. As a narrative genre, sagas provide complications by omitting sometimes a name at all, even if a person is essential in a story (Opheim, 2000).

The *Association for History and Computing* dedicated two special issues of its journal to this theme (in 1992 and 1994). The use of a broad range of software has been reported, varying from regular database management systems, ??e??, to special packages as CARL (Adman, Baskerville and Beedham, 1992), Famlink (Vetter, Gonzalez and Gutmann, 1992) and Genesis (Bloothooft, 1995), and differing strategies have been proposed: rule-based, probabilistic, and learning systems with various degrees of human interaction.

A great deal of the recent discussion has been centred around the so-called 'multi pass algorithms' in automatic record linkage. These are better referred to as subsequently relaxing strategies in comparing nominal records on varying combinations of data elements (e.g., standardised forename and surname, in combination with occupation and year of birth – or Soundexed surname plus shortened forename, combined with occupation, etc.) and the relative confidence attached to these tests. This approach has been introduced by Harvey and Green in identifying voters in Westminster (Harvey and Green, 1994; Harvey, Green and Corfield, 1996), and has been criticised, both on methodological grounds (Adman, 1997) and on basis of manually created true links in nineteenth century census lists (Tilley and French, 1997).

The scope of this section precludes a detailed overview of the rather detailed technical issues involved. More interesting is the overall gain in methodological knowledge. Three aspects are worth mentioning:

- Name standardisation has grown considerably more sophisticated since the early days of the Soundex and Gloria Guth algorithms, as, for example, Bloothooft demonstrated in applying techniques from computer linguistics (Bloothooft, 1994, 1995, 1998).

- Identifying people in a historical context requires an appropriate data structure, which separates original source data from the linkage information. A layered architecture is desirable, where (i) original source data are transformed into (ii) a format suitable for analysis, and to which (iii) the results of the linking process are added as additional knowledge (Bloothooft, 1995; Boonstra and Panhuysen, 1999; Keats-Rohan, 1999; King, 1992).
- This is well illustrated by the COEL system, a digital archive of about 5,000 documents and records, most notably elements from the Doomsday Book, pertaining to the acquisition of English land by the Norman conquerors of the century following 1066. The system comprises three levels. Level one contains all source files, for the most part the text of primary sources, which are given in full. Level two is a database of person names, retaining original Latin forms, and level three represents the interpretative stage, where nominal record linkage has taken place. Here, the user can look for individuals and families, together with commentaries attached. Whatever position in the database the researcher is in, the user is only a double-click away from returning to the original primary source (Keats-Rohan, 1999).[52]
- Different strategies have been tested and compared. However, clear guidelines stating which strategy has to be used with specific kinds of historical data are still missing. That is hardly surprising, because of the variety of conditions that will have influence upon a data set (see above), but would have been in line with the intent of finding generalised linking strategies. Tilley and French, for example, rejected automatic linking by means of multi pass techniques for nineteenth-century census records, but failed to explain in a generalising manner, which source characteristics are most relevant in this respect (Tilley and French, 1997).

3.5.1.7. Stylometry, Content Analysis and Thematics

Knowledge discovery in narrative sources is mainly based on word counting and pattern matching, coupled with advanced statistics. In stylometry[53] and content analysis[54] much work has been hypothesis driven, e.g., about disputed works of Shakespeare, or the authorship of the *Imitatio Christi*, now undisputedly attributed to Thomas a Kempis. There are a few examples of stylometric research in history (Holmes, Gordon and Wilson, 1999; Martindale and

[52] COEL (Continental Origin of English Landholders 1066-1166): http://ahds.ac.uk/creating/case-studies/coel/.
[53] For a short introduction refer to (Rudman, Holmes, Tweedie et al.): http://www.cs.queensu.ca/achallc97/papers/s004.html.
[54] A concise overview of content analysis is available at the website of Colorado State University: http://writing.colostate.edu/references/research/content/index.cfm.

McKenzie, 1995), but most important, in the literary field, is the monumental work by Burrows (Burrows, 2003; Burrows, 1987).

A somewhat related, multifaceted subject is thematics, an interdisciplinary field of study focusing on textual themes. The precise definition depends on the scholarly domain. As Louwerse and Van Peer pointed out in a recent publication, an important feature is that themes allow the 'grouping of meanings into manageable chunks'. Therefore, related information technologies based on thematic analysis can be seen as a form of knowledge management:

> "Themes render the multitude of information meaningful by streamlining individual pieces of information into a meaningful whole which can then be processed more efficiently and linked to ongoing cultural concerns." (Louwerse and Peer, 2002).

Thematic analysis covers many elements in the communicative process, varying from the role of a theme during the process of a text's creation, to the complementary aspect of analysing and modelling how texts are read and understood (cognitive psychology, reception theory, etc.) (Kintsch, 2003; Meister, 2003). The computer program presented by Louwerse is based on the concept of coherence in text; various types of coherence can used in the retrieval of thematic texts (Louwerse, 2003).

3.5.1.8. Conclusions

Both the textual nature of historical and literary sources and the current state of technology are good reasons for discussing text and databases together. In spite of a difference in research objects, historical and literary studies have much in common with regard to computing methods, techniques and tools, e.g., the digitised versions of critical editions, concordances, retrieval facilities and statistical data processing.

Both go beyond the text itself, in literary reception studies, or in mainstream historical studies that use digitised data without aiming at a critical edition. Particularly here, the tension between the abundance of source texts and limited resources of time and money appears, which has given rise to the dichotomy of 'source-oriented' versus 'model-oriented' data processing. Although during the last decade this issue has been amply debated in the community of computer using historians, it seems to be more a practical matter, rather than a dichotomy with a firm methodological basis. The extreme cases will be always clear: when digitising the complete text is an imperative, or when the computer is used to implement a data-driven model. The 'grey' area in the middle is the most interesting and forms a methodological challenge.

The main requirement for any higher level processing is data and text having an appropriate semantic structure. For the time being, creating this structure will require human intervention to a certain degree. However, the challenge is precisely the search for automation. Both large scale historical projects and current developments in other disciplines, like computer science and computer

linguistics, show converging lines into that direction. If scanning is feasible, a wide variety of potentially applicable techniques for further (semi-)automatic structuring do exist. The main problem is not the lack of knowledge, but rather a disparate spread of techniques and expertise over different disciplines, from social sciences, computer linguistics to knowledge engineering and statistics, together with a gap between the theoretical solutions and practical implementations. This is most clearly experienced in the lack of appropriate historical tools.

3.5.2. Statistical methods in historical research

Statistical tools have been applied in historical research for quite a long time. Its use has changed over the years, however. In the 1970s and 1980s, during the heyday of cliometrics, statistics were used predominantly for testing hypotheses, analogous to the way statistics were used in social and economic sciences. Today, statistics are valued much more as a descriptive or even an exploratory tool than as an inductive method, i.e., as a tool that either summarises or helps to find patterns in large historical datasets. The reason for this change first of all has been the concern of many historians that most statistical methods assumed their data to be sampled and distributed statistically in a way to which historical data fell short by definition. A second reason was that historians did not wish to generalise the statistical results they achieved to a larger population; they were satisfied with a mere statistical description of their data, and therefore did not need to test for significant deviations from a null-hypothesis. This does not mean to say that inductive statistics are out of the scope of the historian – on the contrary, a few new inductive statistics have received attention from quantitative history or would be worthwhile for quantitative historians to investigate.

In the following paragraphs a few new statistical tools are described which seem to hold great promise for future historical research.

3.5.2.1. Descriptive and inductive statistics

Logistic regression

Multivariate analysis is of key importance to historians who want to explain variation in a dependent variable by a series of independent variables. Traditional multiple regression techniques are based on a number of assumptions that cannot be met in historical research very easily. Traditional cross tabulation techniques fall short when the number of independent variables is larger than two and interaction effects start to blur the results. Therefore, techniques that are based on fewer assumptions are favoured. At this moment, logistic regression analysis, which has a dichotomous variable as dependent variable, seems to overcome most of the restrictive assumptions traditional (OLS) regression techniques have. In the first place, it does not assume a linear relation-

ship between dependent and independent variables. Secondly, the dependent variable does not need to be normally distributed. Furthermore, normally distributed error terms are not assumed and it does not require the independent variable to be measured at an interval level. Normally, the number of cases in both cells of the dichotomous dependent variable needs to be fairly large. King and Zeng, however, have expanded the use of logistic regression to situations in which the number of cases in one cell of the dependent variable is much less than in the other (King and Zeng, 2001). In doing so, it is possible to start multivariate analysis into the occurrence of rare historical events.

Thus far, the use of logistic regression in historical research has been limited. Only in historical demography and in political historical research analysis, some examples of its use can be traced. In historical demography (Lynch and Greenhouse, 1994) studied the impact of various social, environmental, and demographic factors on infant mortality in nineteenth century Sweden with help of logistic regression; (Reid, 2001) carried out similar methodological work on neonatal mortality and stillbirths in early-twentieth-century England. Derosas used logistic regression to test which variables influenced mobility within the city limits of Venice, 1850-1869 (Derosas, 1999).

In political historical research, logistic regression has been applied to analyse voting behaviour. Schonhardt-Bailey studied voting behaviour in the German Reichstag to test the coalition of landed aristocracy and heavy industry around a policy of tariff protection (Schonhardt-Bailey, 1998); (Cowley and Garry, 1998) tested seven hypotheses of voting behaviour on the Conservative leadership contest of 1990.

Finally, (Henderson, 2000) tried to analyse the extent to which political, economic, and cultural factors are associated with civil wars in sub-Saharan African states, 1950-1992. Results indicated that previous colonial experience was a significant predictor to the likelihood of civil wars. It was also found that economic development reduced the probability of civil war, while militarisation increased it.

Multilevel regression

If one wants to carry out a statistical analysis of historical data, the number of data may be abundant, but the number of variables one can use is rather limited. Especially in micro-level research, where attention is focused on individuals and the relationships they have with their next of kin, the number of variables is small. It would be wonderful if additional data that have been preserved only on other, aggregated, levels could be included into an analysis as well, without losing the statistical tools for testing hypotheses, as is the case in standard regression analysis.

In social sciences, a new technique has become available that indeed is able to cope with different levels of analysis without losing its testing capabilities. This technique, called multilevel regression analysis, was introduced (Bryk and

Raudenbush, 1992). Other, less mathematical, introductions have been published since then (Hox, 2002; Kreft and Leeuw, 1998).

In recent years, multilevel regression has attracted some attention from historians, in particular from a group of historical demographers and economic historians at Lund University in Sweden. For instance, (Bengtsson and Dribbe, 2002) studied the effects of short-term economic stress on fertility in four Swedish parishes, 1766-1865.

Event history analysis

In event history analysis, the behaviour of individuals is studied as a dynamic process of events which did – or did not – take place in the life of the persons under study. The dependent variable in event history analysis expresses the *risk* or *hazard* of experiencing the event. Besides that, event history analysis does not focus on the specific moment in time at which a certain event occurred in the past, but on the time period before it occurred. It does so by looking at the way in which one or more independent variables may have contributed to the relative risk of the event to happen. It is important to realise that, because event history analysis measures periods of time and not points in time, data can be appended to the analysis which are only available for a limited time span. Even if we do not know what happened to a person before or after such a time span, the information that is contained in this particular time span still can be put to use. It is this inclusion of so-called 'censored' data, which makes event history analysis a very powerful tool for analysing dynamic historical data.

Good introductions to event history are (Allison, 1984) and (Yamaguchi, 1991). A good manual which can serve both as a student textbook and as a reference guide for the use of event history analysis is (Blossfeld and Rohwer, 2002). There is a software program called TDA (Transitional Data Analysis), which is written exclusively for event history analysis.

Although originating in biomedical research (where it is called survival data analysis), event history analysis has been welcomed by historians in a pretty early stage, for instance by (Raffalovich and Knoke, 1983). Especially in historical demography, the technique has become a standard procedure (for an overview see (Alter, 1998). Already in 1984, the first applications in historical demography have been published (Egger and Willigan, 1984). In recent years, historical demographic studies have been done on child mortality by (Bengtsson, 1999), on mortality by (Campbell and Lee, 1996) and (Bengtsson and Lindström, 2003), on child bearing by (South, 1999), and on marriage by (Gutmann and Alter, 1993) and (Cartwright, 2000).

Social historical research based on event history techniques have been done on social issues like migration by (Schor, 1996), (Kok, 1997), (Campbell and Lee, 2001), on legislation by (McCammon, 1999) and on inheritance by (Diekmann and Engelhardt, 1999).

Finally, event history analysis has been applied to the field of economic history, notably on employment issues (Alter and Gutmann, 1999; Drobnic, Blossfeld and Rohwer, 1999).

Although event history analysis is widely used, there are some unresolved problems attached to it. Some have to do with the assumptions that accompany various different event history models, while other problems have to do with the impact censored data still have on the results, or with the interpretation of results when an "event" has more than one meaning. In any case, event history analysis has proven to give new insight in historical processes.

Ecological inference

Because of the lack of data on the individual level, historians often use aggregated data in order to gain more insight into individual's behaviour. In doing so, there is always the possibility of falling into the trap of 'ecological fallacy', i.e., the problem that arises from attempts to predict individual behaviour based on aggregate data of group behaviour.

In the past a few attempts have been made to solve this so-called "ecological inference" problem. Goodman proposed a solution in 1959, called ecological regression (Goodman, 1959). Although this regression technique was able to overcome some of the ecological inference problems, it created another: the results of an ecological regression were hard to interpret. For instance, standardised regression coefficients well over the maximum limit of 1 often appeared as a result.

In 1997, Gary King presented a different approach to the problem. "A Solution to The Ecological Inference Problem: Reconstructing Individual Behaviour From Aggregate Data" was a book that introduced a new statistical tool; EI and EzI were the accompanying freeware computer software programs (King, 1997).

Historians responded quickly to this new approach. In 2001, *Historical Methods* published two volumes in which King's solution was introduced and evaluated (starting with an introduction by (Kousser, 2001)). This was done by replicating historical research with known individual data on an aggregate level with King's (and Goodman's) method to see whether the results would coincide. On the whole, the conclusion was that King's method might be *a* solution, but definitively not *the* solution. Nevertheless, King's method has been embraced by other sciences, especially social and political sciences, and it is to be expected that it will be used in historical research as well.

Time series analysis

An important category of techniques for historical research is time series analysis. Time series models can be applied to all sources for which diachronic series exist, not only economic sources such as trade statistics or shipping movements, but also to demographic sources such as population data. Therefore, time series analysis can be applied to a great variety of historical research top-

ics. However, applications outside the domain of economic history are remarkably scarce (Doorn and Lindblad, 1990).

In population studies, three types of time-series effects are distinguished, having to do with age, period, and cohort. Age effects are effects related to ageing or the life cycle. For instance, individuals often tend to become more conservative as they age. Period effects are effects affecting all cohorts in a given historical period. For instance, individuals who experienced the Great Depression became more likely to support social welfare policies. Cohort effects are effects which reflect the unique reaction of a cohort to an historical event, or which were experienced uniquely by the cohort. For instance, the post-WWII cohort, reaching draft age during the Vietnam War, experienced unique issues that seem to be associated with increased alienation from government. Disentangling these three types of effects for one set of time-series data is a major challenge of time series analysis.

Some examples can be found in historical demography (for instance on mortality by (Bengtsson and Broström, 1997) and on illegitimate fertility decline in England, 1850-1911 by (Schellekens, 1995). Other examples come from the field of political history, where (Pacek and Radcliff, 2003) tested whether Left-wing parties were the primary beneficiaries of higher rates of voter turnout at the poll-box. Time series have been applied in historical climate research, where Schableger analysed time series for the daily air temperature in Vienna between 1874 and 1993. Climate research was also part of a spectacular interdisciplinary time series analysis by (Scott, Duncan and Duncan, 1998) on four centuries of various grain prices in England (1450-1812). Their analysis revealed cyclic effects of changes in weather conditions.

Finally, there are some examples of applying time series in social history. Although a nice introduction of time series to social history has been published by (Stier, 1989), only very few social history articles have been published thus far. The papers by (Isaac, Christiansen, Miller et al., 1998) on the relationship between civil rights movement street tactics and labour movement militancy in the post-war United States, and (Velden and Doorn, 2001) on strikes in the Netherlands are an exception to the rule.

In economic history, things are different. In this domain, time series analysis has remained alive and well over the past decades, and applications have been diverse. A few special applications stand out. First of all, time series analysis was used for international comparisons. Raffalovich investigated the impact of industrialisation, economic growth, and the unemployment rate on property-income shares in a sample of 21 nations in Europe, Asia, North America, and the Pacific area during 1960-90 (Raffalovich, 1999). Li and Reuveny analysed the effect of globalisation on national democratic governance 1970-1996 for 127 countries in a pooled time-series, cross-sectional statistical model (Li and Reuveny, 2003). Finally, (Greasley and Oxley, 1998) reviewed three approaches to explain the timing, extent and nature of shifts in British and Ameri-

can economic leadership since 1860. They concluded that better educational opportunities played an important role for the U.S.A. to gain economic leadership in the twentieth century.

The relationship between public expenditure on education and economic growth has also been studied by (Diebolt and Litago, 1997) for Germany and by (Ljungberg, 2002) for Sweden.

Finally, there has been methodological interest in time series analysis as well. There was a plea for time series as a descriptive tool instead of a tool to test hypotheses by (Metz, 1988a). New tools for time series analysis were introduced: for instance, methods for the analysis of long-term cycles like the Kondratieff cycle (Diebolt and Guiraud, 2000; Metz, 1988b, 1993). Another example is the application of new methods of filtering in order either to discern various cycles from one another (Mueller-Benedict, 2000) or to analyse non-stationary stochastic processes (Darné and Diebolt, 2000).

3.5.2.2. Exploratory data analysis and data mining

Data mining is a general term for a variety of techniques that are meant to gain insight into a data set. Most often, its goal is to uncover the underlying structure of the data set, but it may also have the purpose to find out what variables are the most important ones, or what cases are connected to each other or what cases are the outliers in the dataset.

If the techniques employed are based on statistical formulas, data mining is called "exploratory data analysis". It is these techniques that are presented here. The way to present results from exploratory data analysis can take many shapes, but often a graphical representation is favoured. Most of such visual methods are described in the next paragraph on visual data analysis. In this paragraph only two methods of exploratory data analysis are presented, cluster analysis and simulation.

Clustering techniques

Cluster analysis seeks to identify homogenous subgroups of cases in a population. That is, cluster analysis seeks to identify a set of groups in which within-group variation is minimised and between-group variation is maximised. There are many different clustering techniques, many of which are hierarchical. In hierarchical analysis, the first step is the establishment of a similarity or distance matrix. This matrix is a table in which both the rows and columns are the units of analysis and the cell entries are a measure of similarity or distance for any pair of cases. The second step is the selection of a procedure for determining how many clusters are to be created, and how the calculations are done. The third step is to select a way to present visually the results of a cluster analysis. Normally, a (cluster-) dendrogram is the method that is used most often, but historians also use geographic information systems to plot the results. Finally, statistical analysis on the results of cluster analysis can help to interpret the outcome.

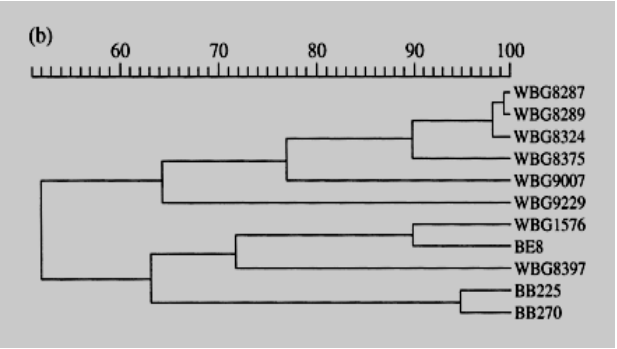


Figure 3.7. A dendrogram, showing, among other things, the similarity between WBG8287 and WBG8289, and the dissimilarity between the first six cases and the remaining five. From (Price, O'Brien, Shelton et al., 1999).

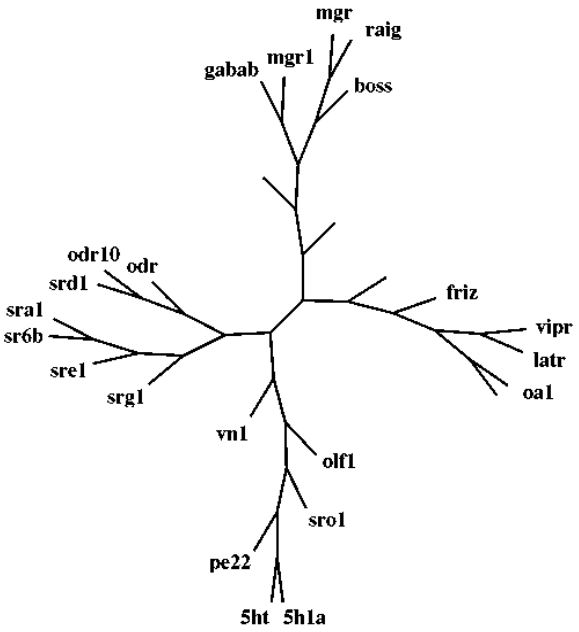


Figure 3.8. A cluster-dendrogram, showing four groups of cases. From (Graul and Sadée).

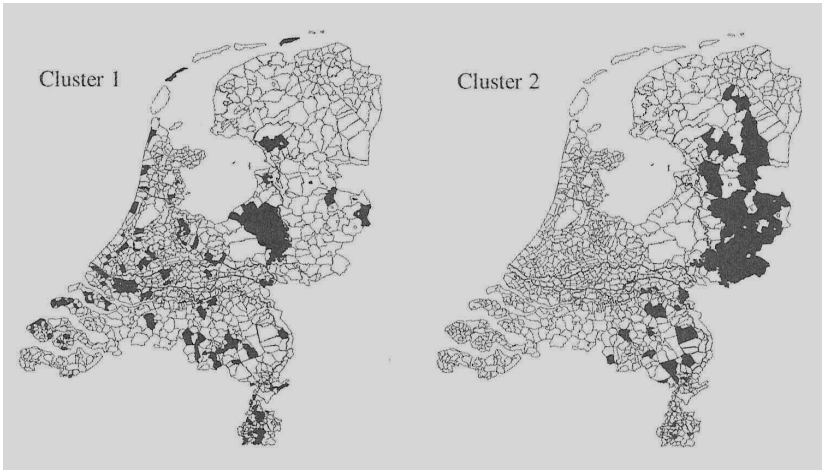


Figure 3.9. Geographical representation of two clusters of municipalities, The Netherlands 1899. From (Boonstra, 2001).

Around 1990, a few introductions to cluster analysis in historical research were published (Bacher, 1989; Boonstra, Doorn and Hendrickx, 1990). Since then, cluster analysis has been used by a variety of historians for exploratory purposes. Cluster analyses, in which geographical information systems have been used to visualise results, have recently been done on the basis of nations (Obinger and Wagschal, 2001), provinces (Delger, 2003), districts (Debuisson, 2001), municipalities (Boonstra, 2001), and parishes (Song, 2002). Cluster analysis on households has been done by (Galt, 1986) and (Spree, 1997).

Simulation

Although the idea of "simulation" is simple enough, the statistical basis of its scientific use is very diverse and often very complex. Computer simulations are designed to evaluate behaviour within a specifically defined system, establish models and criteria, develop and analyse strategy, predict outcome, determine probability, and identify relational patterns.

A model is always the starting point of simulation, and a correct operationalisation of all variables and links between variables in the model is a prerequisite for good analysis. When the model has been established, a series of simulation runs can be done. During these runs, various techniques can be applied. One way of simulating is by changing variable parameters, so that the impact of these changes on the model can be measured. When the simulation process consists of a finite number of states, in which the future behaviour of the system depends only on the current state and not on any of the previous states, this is called a "Markov chain" model.

Another way of simulating is by repeating the same model with the same parameters over and over again in order to find out how stable the model is through all of these runs. Repeated simulation can also be used to create a hypothetical dataset, which can then be tested with various statistical tools. Simulations of this kind are called "Monte Carlo" experiments.

Simulation techniques were widely established in the social and economic sciences in the 1970s. In historical research with its wonderful possibility to check historical simulation results with the outcome of real historical events and developments, applications have been rather scarce, however. The use of simulation techniques, especially in political history has been disappointingly low, although its importance has been stressed (Mielants and Mielants, 1997). A very early exception to the rule has been the semi-computerized attempt to simulate the outbreak of Word War I ((Hermann and Hermann, 1967). A more recent example is the research done by (Artzrouni and Komlos, 1996), who devised a very simple simulation model in order to find out why Western European states achieved stable boundaries much earlier than Eastern European states. The model does not take many factors, such as economic differences, state policy, or military effectiveness, into account. Nevertheless, it demonstrates that geographical constraints have played an important role in determining the map of modern Europe. A final political example is from (Deng, 1997), who made a simulation model of the 1989 Chinese student movement. From the model it becomes clear that the Chinese government, by concealing its preferences, caused the death of many demonstrators who believed that the army would never harm the Chinese people. The model also makes clear that an information gap can lead to unintended and undesirable outcomes, even when actors behave rationally.

Most applications can be found in economic history and historical demography. In historical demography, an important impetus to the use of simulation was given in Reher and Schofield's book on new methods in historical demography (Reher and Schofield, 1993). In this book, a special part was reserved for research in which historical demographic processes were simulated. More recent studies, in which simulation techniques are used for historical demographic research, are for instance (Brod, 1998), who used Monte Carlo simulation for the analysis of marriage seasonality. Zhao investigated the relationship between demographic conditions and family or kin support systems in Victorian England with help of a simulation model in which kinship patterns change during the life course (Zhao, 1996). Hayami and Kuroso used a similar approach for their research into the relationships between demographic and family patters in pre-industrial Japan (Hayami and Kurosu, 2001). Okun used simulation techniques in order to distinguish between stopping and spacing behaviour in historical populations (Okun, 1995). In doing so, she was able to determine what the principle method of regulating family size was during the demographic fertility transition of the nineteenth century. McIntosh used a

simulation model in trying to solve the question why populations of small towns in southern Germany stagnated following the Thirty Years War (McIntosh, 2001).

Simulation models in economic history have been used to study macro-economic effects in nations like Germany (Ritschl, 1998) and Russia (Allen, 1998), urban systems (Guérain-Pace and Lesage, 2001), or in periods of time like the Industrial Revolution (Komlos and Artzrouni, 1994). At the micro level, the work of the Computing & History Group at Moscow State University must be mentioned. Andreev, Borodkin and Levandovskii, all members of the group, used simulation models for an explanation of worker's strikes in Russia at the beginning of the twentieth century (Andreev, Borodkin and Levandovskii, 1997). One of these models, using chaos theory as a starting point, even pointed towards a 'spike' in strike dynamics for 1905, even though there was no input about events relating to the Revolution of 1905. Borodkin and Svishchev used a simulation model based on Markov chains to find out more about the social mobility of private owners under the New Economic Policy (NEP) in the Soviet Union during the 1920s (Borodkin and Svishchev, 1992).

Finally, simulation techniques have also been applied in historical environmental research. Nibbering and De Graaff developed a watershed model for an area on the island of Java, using historical data on land use in order to simulate past hydrological conditions and erosion (Nibbering and DeGraaff, 1998). Allen and Keay analysed various simulation models to find out what caused the bowhead whale to be hunted almost to the point of extinction by 1828 (Allen and Keay, 2001).

3.5.2.3. Texts and statistics

As mentioned above, stylometry, authorship attribution and content analysis are important techniques for knowledge discovery. This does not mean to say that there is *communis opinio* about tools for authorship attribution. On the contrary, recently quite a few new methods have been introduced. In all introductions, the *Federalist Papers* serve as a benchmark for testing various tools. The 85 *Federalist Papers* were written by someone called 'Publius' in 1787 and 1788 to persuade the citizens of New York State to adopt the nascent Constitution of the United States. But 'Publius' was not a single person; of the 85 texts, 52 were written by Alexander Hamilton, 14 by James Madison, 4 by John Jay, and 3 by Hamilton and Madison jointly. The authorship of the remaining twelve texts is disputed.

The Federalist Papers were used for authorship attribution for the first time by (Mosteller and Wallace, 1964). Thirty years later (Holmes and Forsyth, 1995) came up with a few new solutions, which triggered off new research by many others. Tweedy et al. applied a neural network analysis (Tweedie, Singh and Holmes, 1996), (Baayen, van Halteren and Tweedie, 1996) proved that attribution results could be improved by annotating the texts with syntactic

tags, whereas Khmelev (Khmelev, 2000; Khmelev and Tweedie, 2001) showed that the success rate would be even better using Markov chain analysis on letters.

Against impressive results there are also principal objections in post-modern discussions about authorship. A much-voiced criticism boils down to the argument that stylometry has indeed speed and power, but that it doesn't know what it is counting (Merriam, 2002).

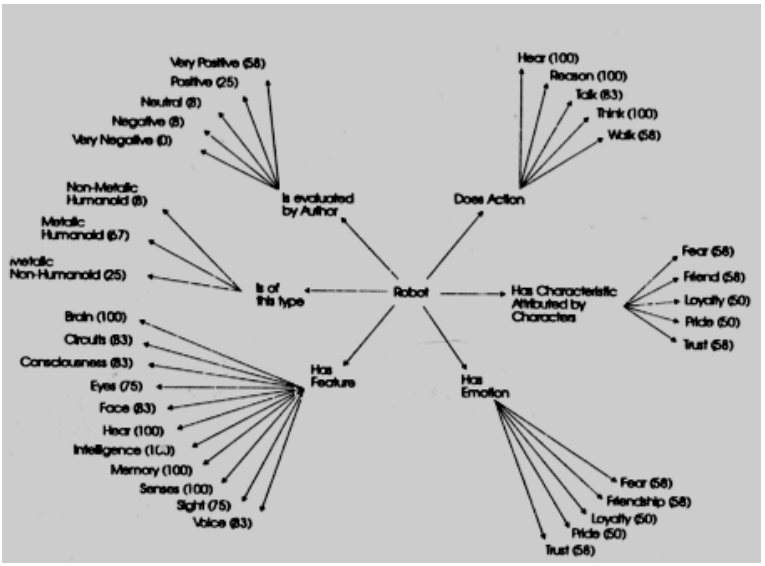


Figure 3.10. Content analysis: A map representing relationships among concepts in robot descriptions. From (Palmquist, Carley and Dale, 1997)

Content analysis dates also back to the 1960s, when the famous Harvard program *The General Inquirer* was deployed in automatic analysis of political documents, folk tales and private correspondence, revealing secrets that could not be caught by the naked eye. A more recent example is a study about the changing depiction of robots in writing over more than a century. Content analysis has been deeply rooted in the social sciences, and has not received the attention it deserves in historical research.[55] In some cases stylometric research comes close to content analysis, as, for example, with the multivariate analysis of two texts of the American novelist Charles Brockden Brown. Analysing the

[55] For example some projects at the Zentrum für historische Sozialforschung (Cologne): http://www.za.uni-koeln.de/research/zhsf/, the analysis of the historical newspaper corpus at Rostock (http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/llc/Conference1998/Papers/Schneider.htm) and (Breure, 1999).

novels *Wieland* and *Carwin* Stewart was able to show that Brockden has succeeded in creating a narrator with a distinctive voice, thus providing evidence which could not be obtained through normal reading and aesthetic interpretation (Stewart, 2003).

Both, stylometry and content analysis provide additional structure by adding a connection between text parts and a conceptual level (e.g. a list of topics, events or authors). Once established, these links can be utilised in search and retrieval operations.

3.5.2.4. Conclusions

Recently, quite a few new statistical techniques have been developed that hold a great promise for historical research. Although varied in the statistical results aimed at, as well as in the underlying mathematical formulas, they are promising, because they possess at least one of the two characteristics that have been described above: they have a much better fit with historical data, and they are much more in line with the traditional methodology of historical science. Logistic regression, multilevel regression, event history analysis and ecological inference are examples of techniques with such a better fit; the various techniques for exploratory data analysis are examples of a better fit with the traditional research methodology of historical science.

What will be most interesting, will be the development of new statistical techniques that possess both characteristics. Traditionally, there has been already such a technique: cluster analysis. Its use, however, has been hampered by the fact that in social science, from which most statistical tools are derived, researchers disapprove of cluster analysis because of its inability to test hypotheses. Therefore, it is to be expected that new methods for exploratory data analysis will not be developed by social scientists. It will be information scientists who will do the job, especially within the framework of data mining research. Furthermore, it will be information science that will come up with new methods to present visual results from data mining. It is this combination, in which visual presentation tools are added to techniques for extracting information from large datasets that will set the agenda for research into statistical methods for historical science in the near future.

3.5.3. Images: from creation to visualisation

Until very recently, serious historians thought images to be of limited value to historical research. Of course, in popular historical writing, images always have played a pivotal role; very often, text in such histories was of minor importance. It was the images, and their captions, that drew the reader's attention first. In serious, scientific, historical writings, it was the other way around: images were only valued when they could illustrate a historian's text – they played their part when words fell short.

At the moment, this situation is changing rapidly. First of all, the cost of displaying images in print, which was a serious impediment to using images more extensively, has diminished. Especially in the new forms of presentation that have come about, like Internet publishing, costs have stopped being a problem at all. But the reduction in costs is only a small part of the explanation for the growing use of images in historical research. Another, much more important reason is the developing interest in the 'culture of images' and in the image as a historical source (Burke, 2001; Zeldenrust, 2003). There are even attempts to "write" a history by showing pictures only (Staley, 1998, 2003). For a long time, research on images has been the exclusive domain of cultural history, but nowadays interest has awakened by social and economic historians as well.[56]

Additionally, the amount of historical images available for research has grown tremendously over the past few years. More and more images have been digitised and put into visual archives, opening up a source for research by allowing historians much easier access to thousands of images than ever before. "Images" have become "information", and fit into the life cycle of historical information in the same way as textual or numerical data. As a consequence, the possibilities for historical analysis on the basis of images have grown as rapidly, and, as a consequence, the methodological implications for using images in historical research have become an issue.

In this section we will deal with these methodological issues at the various stages in the lifecycle of digitised historical images, from the phase of creation, via enrichment, retrieval and analysis to the phase of presentation. To do so, we pay attention mainly to digitised historical photographs. But there are other kinds of digital images than historical photographs, e.g., movies and graphics, etc. which all have different characteristics for storage, analysis or presentation. Whenever appropriate, these other types will receive special attention in this section as well. The section ends with two special paragraphs: one devoted to visualisation of textual data, and one devoted to historical geographic information systems, and the maps these systems can produce.

3.5.3.1. Creation

If one wants to use thousands of images in order to do research into a specific historical period of time or into a specific theme, one cannot proceed without digitising the images and putting them into a database. At the moment, already hundreds of databases of this kind have been set up, many of them accessible through the Internet. They vary greatly in size, from hundreds into millions of images.

[56] See for instance the international symposium on History and New Media, which was held in Berlin, 9-11 April 2003. More information on: http://clio-online.de.

There are a few introductions to digitising photographs. A book by (Frey and Reilly, 1999) contains detailed discussions of technical and quality issues with lots of illustrations. Ostrow is one of the few who published a guide for digitising historical photographs for the internet (Ostrow, 1998).

This does not mean to say that there is a standard way of creating digitised images. Already this first step in the lifecycle of digitised images shows a wide range of solutions. Some databases contain only low-resolution or compressed images, in order to reduce costs or to overcome copyright problems, while other databases contain different copies of one image, each with a different resolution. Some databases contain images that have been enhanced with special software in order to improve the pictorial quality of the image; others have explicitly decided not to do so.

Finally, image databases also differ in the way the images are stored and preserved. Preservation techniques are object of numerous studies, for instance within the framework of the European SEPIA project.[57] Klijn and De Lusenet have presented an overview of various datasets that have recently been set up in the E.U. (Klijn and Lusenet, 2002). The report shows that there is no standard in the technical process of digitising, no standard in the quality of the digitised images, no standard in the preservation techniques employed and no standard in the use of metadata to describe an image.

3.5.3.2. Enrichment

A very important issue is the enrichment of digitised photographs in order to be able to retrieve them systematically and comprehensively. There are two major approaches to image information retrieval: content-based and metadata-based image retrieval. In the metadata-based approach, image retrieval is based on descriptions of images. The problems surrounding what to describe and what not are numerous. Until recently, databases of photographic collections were set up mainly by art historians. The metadata they used mirrored their interests, describing for instance in detail the kind of paper the original photograph was printed on, leaving a description of what was on the picture aside.

But making a description of what can be seen on a picture is easier said than done. Contrary to other kinds of data, a description about what is on a picture can take many forms, and will be different for almost every single viewer. The picture in Figure 3.11, for instance, can serve as an example (Klijn and Sesink, 2003).

[57] SEPIA project: Safeguarding European Photographic Images for Access. http://www.knaw.nl/ecpa/sepia/.

Figure 3.11. Information about a picture's content can be different for various viewers: does the picture show the KNAW main building, a Volvo 234 turbo fuel injection, or autumn on a canal in Amsterdam?

Nevertheless, at the moment, a lot of effort is being made towards a standardised description of photographic images. From the European SEPIA project, a proposal has been made to have a picture described with no less than 21 fields of information (Klijn and Sesink, 2003).

Main reference code	Geographical location
Name of institute	Access restrictions / copyright
Acquisition code	Relationships
Location	Status
Description	Technical identification Dimensions
Title	Photographic type
Creator	File format
Descriptors / subject headings / classification	References
	Origins of collection / grouping
Names	Contents of the collection / grouping / acquisition
Date	

Figure 3.12. SEPIADES' 21 core elements to describe a photograph

The metadata-based enrichment of video images poses a special problem, because the image that needs to be described changes constantly. A video therefore needs to be divided into segments in order to describe them sepa-

rately. To do this efficiently, automated methods for segmenting videos are called for. At the moment, some effort is being put into such methods within information science: a few parsers have been developed which divide a video stream into constituent building blocks (Koprinska and Carrato, 2001). These efforts primarily have the intention to compress video data files, but can be used for classification purposes as well.

3.5.3.3. Retrieval

As has been stated before, there are two kinds of retrieval systems: content-based and metadata-based retrieval systems. In content-based image retrieval the images are retrieved based on the characteristics of the digitised image, such as colour, texture, shape, etc. Its advantages are obvious: there is no need for adding metadata to the database, keeping the cost of a digitised collection of images very low. Content-based image retrieval is a hot issue in information science at the moment. At the moment, work is done on dozens of content-based image retrieval systems, an overview of which is given by (Gevers and Smeulders, 2004). Some of the systems use historical photographs to test the quality of the system.[58] Although fascinating the results suggest that the prospects of content-based image retrieval systems for historical research are not good.

Therefore, historians and archivists have pinned their hopes on metadata-based retrieval systems. To retrieve images from a database, the queries normally are based on keywords. Because of the lack of standardisation at the enrichment of images, the precision and recall of such search methods is often disappointingly low. It is therefore better to use ontology-based annotations and information retrieval (Schreiber, Dubbeldam, Wielemaker et al., 2001).

A solution to the problems related with content-based and metadata-based retrieval systems could very well be a combination of both. A prototype of such a system, called "Ontogator" has been put to use with a photo repository of the Helsinki University Museum (Hyvönen, Saarela and Viljanen, 2003).

A second issue of interest to historians and information scientists, is the creation of ontologies for various image collections, or systems to query various data sets. An interesting example, especially for historians, is the German Prometheus project (Hartmann), which is based on the ??e?? database software. Image databases are located at various servers; a central server only works as a broker between these image databases and the end-user, making the impression to the end-user that it makes his queries and views the results from one single database. Prometheus also includes a number of new devices to retrieve and analyse digitised images.

[58] For instance, in Leiden, where a content based image retrieval system is being tested on a database of 21,094 studio portraits, 1860-1914. http://nies.liacs.nl:1860.

3.5.3.4. Analysis: visualisation

Visualisation, or visual data analysis, refers to computerised methods in which data are synthesised and summarised in a visual way, comparable to the way statistics recapitulate data in a formal numerical way. In this sense, visualisation is not meant to make a presentation of known (statistical) results as is done with traditional graphics and maps, but is meant to explore data in order to generate ideas and explore alternatives (Grinstein and Ward, 2002; Gröller, 2001). In so doing, visualisation offers a method for seeing the unseen.

The interest in visual data analysis has grown immensely during the last couple of years in many sciences, especially information science. In historical research, examples of its use are hard to find, except for the use of cluster analysis (see Section 3.5.2).

Like cluster analysis, visual data analysis falls within the framework of explorative data analysis, as a means of gaining understanding and insight into the data in a visual way. The difference between explorative visual data analysis and visual presentation is small, however. Similar tools can be used in both ways: if a special technique allows researchers to explore and interpret their data well, it will serve as a means to present their data efficiently as well.

Visual exploration of data can take many forms, for instance (Gershon and Eick, 1997; McCormick, Defanti and Brown, 1987; Reinders, 2001). For historical research, the visual analysis of time-varying data will be most interesting. A number of applications have already been developed over the years, even in the pre-computer era (Tufte, 1983).[59] A few examples of new visualisation tools are given below.

[59] A nice overview of classic contributions to visualization can be found at: http://www.csiss.org/classics/.

Lifelines

Among the first computerised visual data analysis applications in history has been the display of the dynamics of household composition, originally created by the Swedish Demographic Database at Umeå in de 1970s (Janssens, 1989). Time runs from left to right, starting with the moment of household formation. Children enter the graph when they are born and leave the graph when they die or migrate from the household. The same procedure is kept for parents, grand-parents, other relatives and lodgers. In this visual way, a comparison can be made between the dynamics of various households.

A very similar approached has been use by (Plaisant, Milash, Rose et al., 1996) in order to visualise personal histories.

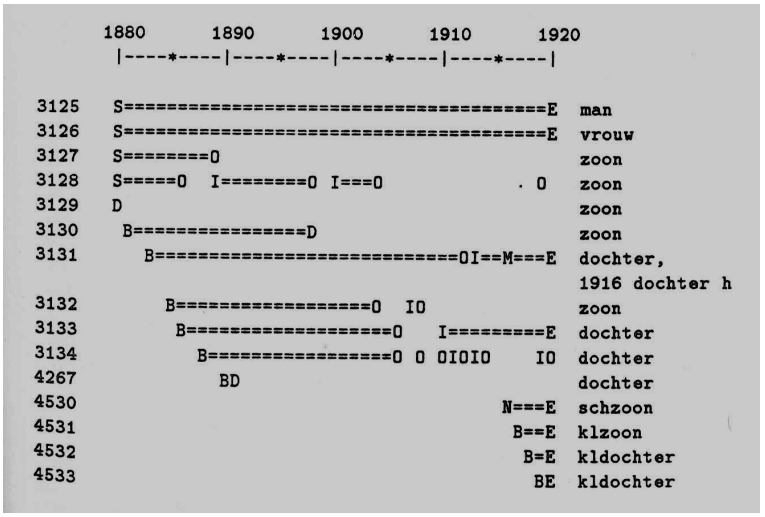


Figure 3.13. Lifeline of a household, 1880-1920 (Janssens, 1989). Explanation of symbols used: S start of observation, E end of observation, B birth, D death, O migration out, I migration in, M marriage, N entry.

The Lexis pencil

There has been little development of graphical methods to visualise an event history dataset. This is partly because of the complexity of such datasets; it is easy to be overwhelmed with the number of variables and different dimensions of a typical study.

In order to display event history data, a Lexis "pencil" (introduced by (Keiding, 1990)) can be made for each case history. Three variables can be selected for three co-ordinate axes. Normally the Y-axis at least will be assigned a variable that is some measure of time. Any of the remaining variables may be used to "paint" the faces of the Lexis pencils with colours according to the value of the variable. There will normally be one variable that is an index of the case histories. The resulting "geometry" is displayed in a window together with a set of axes and annotation.

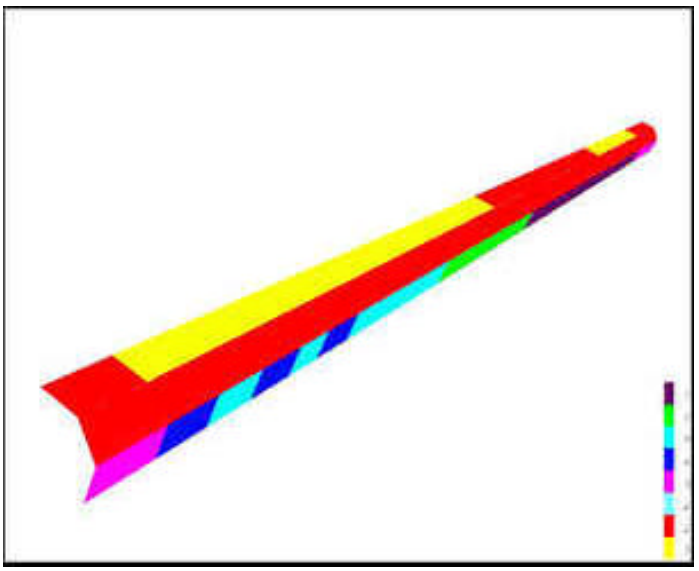


Figure 3.14. The life of a married couple displayed as a Lexis pencil. Time runs from left to right, starting at date of marriage and finishing at the survey date. Each face of the pencil represents a different variable. The top face represents the employment history of the wife, the middle face that of the husband, and the bottom face the age of the youngest child in the household. From (Francis and Pritchard, 1998).

Calendar view

Van Wijk and Van Selow use time series data to summarise them in a calendar-like fashion in order to find changes and irregularities in standard patterns (Wijk and Selow, 1999). A similar approach, called "the agenda", has been proposed by (Daassi, Dumas, Fauvet et al., 2000).

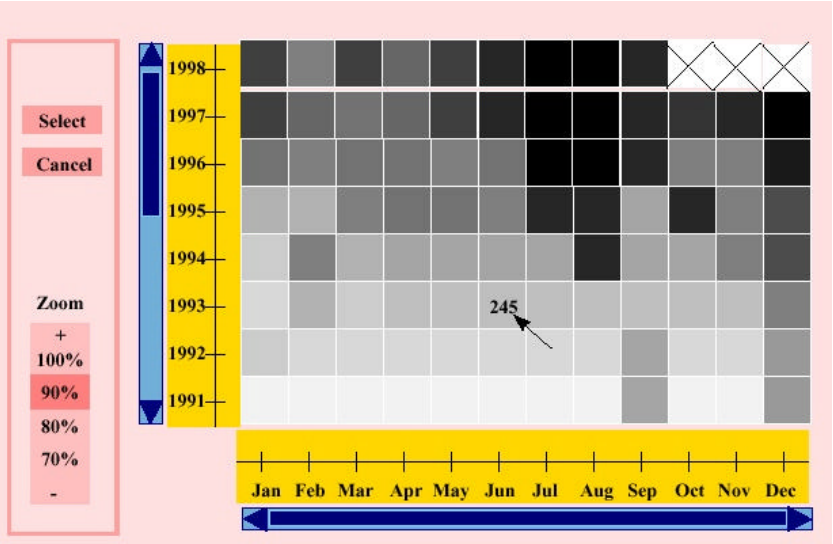


Figure 3.15. "The Agenda". Visualisation of the monthly productions of an assembly line using the calendar technique (Daassi, Dumas, Fauvet et al., 2000).

Concentric Circles

Concentric rings represent different variables. The colour, shading and width of the rings as they are traversed in a clockwise direction represent changes of the variables over time (see (Barry, Walby and Francis, 1990); see also (Daassi, Dumas, Fauvet et al., 2000))

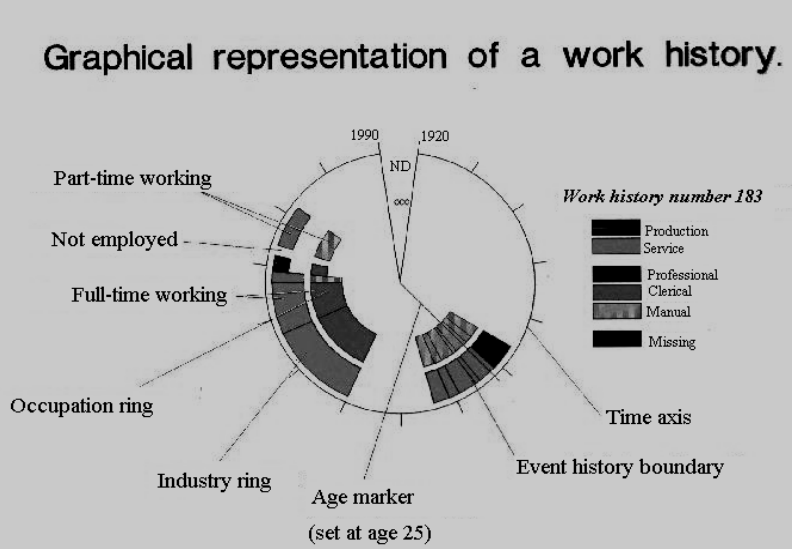


Figure 3.16. Example of the concentric circles technique, illustrating a lifetime work history. Colour is used to represent movement in and out of different industrial sectors and social class; the width of the rings represents the number of hours worked. From (Barry, Walby and Francis, 1990).

A number of more or less similar visual techniques have been developed in order to find irregularities in time series (Keogh, Lonardi and Chiu, 2002).

3.5.3.5. Visual analysis of texts

A novel application of visualisation techniques is in the exploration of texts, especially in large textual datasets, where visual interfaces to digital libraries apply powerful data analysis and information visualisation techniques to generate visualisations of large document sets (Börner and Chen, 2002). "Semantic timelines" may serve as an interesting example for historians, because it includes time as a classifying principle for historical texts. Since much historical information involves commentary on previous data, different time periods are vertically stacked, so that a period of comment or consequences can sit directly on top of a period of earlier activity. In addition, Semantic timelines are zoom-

able, can illustrate time granularity and indeterminacy, and can include hierarchical nesting structures for drill-down. Together, these techniques allow a vast amount of historical information to be viewed within a rich semantic context, at various scales simultaneously (Jensen, 2003).

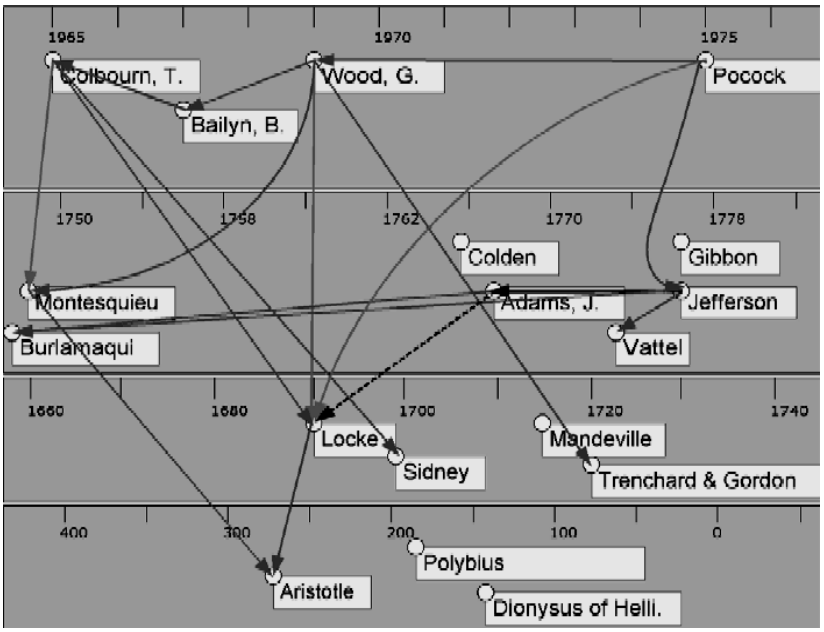


Figure 3.17. Four time periods are vertically stacked, showing references among publications dealing with republican theory. Arrows indicate support, opposition, comment and familiarity with the issue at stake (Jensen, 2003).

A second example is "AuthorLink", a project by Jan Buzydlowski. AuthorLink explores author relationships through co-citation patterns. The assumption is that if two authors are often cited together by many other authors, these two authors likely have a common intellectual interest in their research and writing. When many related authors' pair-wise co-citation patterns are explored, we will have a map of a subject domain where authors on the map represent ideas or subtopics as well as their relationships (Buzydlowski, White and Lin, 2002).

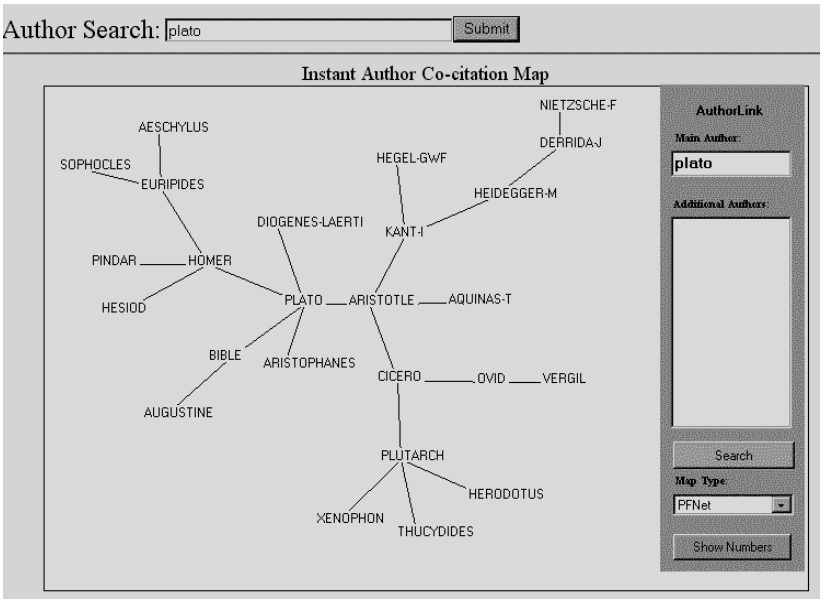


Figure 3.18. Example of the use of AuthorLink. The text base used by AuthorLink is the Arts & Humanities Citation Index, 1988-1997, comprising a total of about 1.26 million records. A search was done on Plato as the main author. From (Buzydlowski, White and Lin, 2002).

In the Netherlands, an approach similar to AuthorLink has been put into practice within the historical domain by the *Digitaal Erfgoed Nederland* consortium. They use a commercial product called "Aquabrowser" to find co-occurrences of words in various websites.[60]

But textual databases of lesser size can benefit from visual data analysis as well. Monroy et al., for instance (Monroy, Kochumman, Furuta et al., 2002a; Monroy, Kochumman, Furuta et al., 2002b), developed a way to study differences in various early text editions of Cervantes' Don Quixote with visualisation tools.

[60] De Cultuurgrazer: http://den.medialab.nl/.

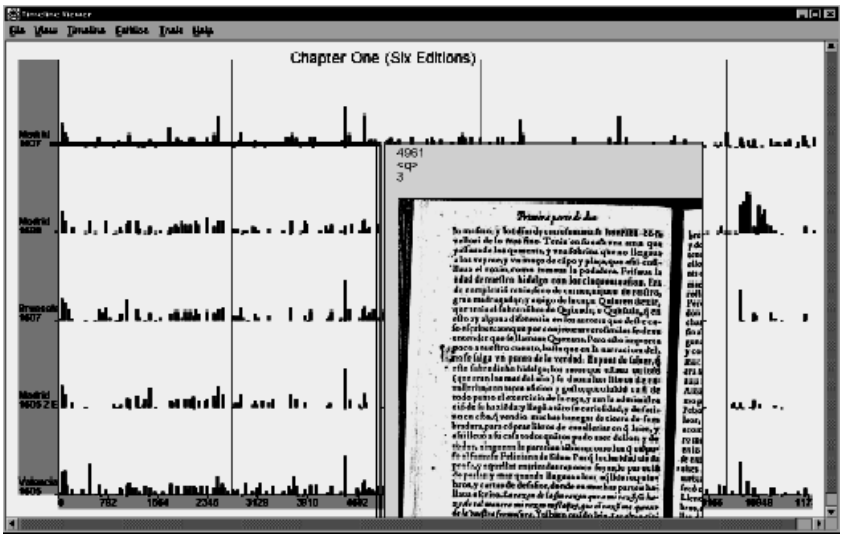


Figure 3.19. A timeline viewer, depicting variants among six early editions of Cervantes' Don Quixote. (Monroy, Kochumman, Furuta et al., 2002b).

Finally, Lecolinet et al., who introduced tools for automatic line segmentation of scanned hand-written manuscripts (see also Section 3.5.1.5.), also showed innovative methods for visualising literary corpus which contains many versions or variants of the same manuscript pages (Lecolinet, Robert and Role, 2002).

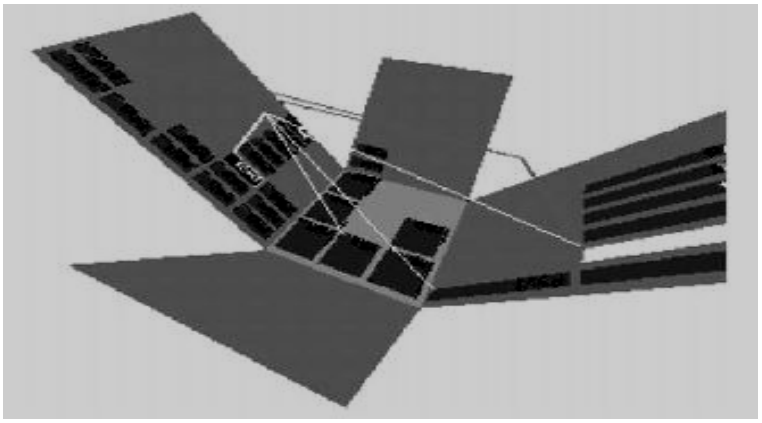


Figure 3.20. Perspective Wall model adapted for visualising manuscript collections. From (Lecolinet, Robert and Role, 2002).

It is to be expected that visualisation of historical data will become a tool to take off in the near future in historical science. However, it will take a while before application will be wide spread. Before that, standards need to be developed: standards on how to visualise certain kinds of data, and, especially, standards on how to interpret visual representations (Chen and Börner., 2002).

3.5.3.5. Historical GIS

> "It is easy to predict that when we recollect the development of history at the end of the twentieth and the beginning of the twenty-first century, the introduction of GIS to research and teaching about the past will be one of the signs of the successful continuation, and reinvigoration, of that tradition [of innovation in historical research]" (Guttman, 2002)

If we define geography as the study of spatial differentiation, and history as the study of temporal differentiation, historical GIS can be defined as the study spatial patterns of change over time (Knowles, 2002).

"GIS" refers to "Geographic Information System", an integrated system in which geographic co-ordinates and research data are stored, as well as tools retrieve and analyse information.

Until recently, historians have made only limited use of GIS tools. There are a number of reasons why this has been the case. First of all, historians for a long time have thought of maps only as a means for presenting data instead of analysing them. For such a limited scope, GIS software was too expensive to use for a long time. Next to that, the data structure of GIS software was very exotic, while at the same time the software was not very capable of importing data into its system. And in the third place, standard GIS software seemed poorly suited to handle geographic changes over time.

Nevertheless, an overview in 1994 revealed that there were a number projects being carried out in various countries spread across Europe and the United States (Goerke, 1994). But that number is negligible if look at the use of historical GIS nowadays. And rightly so: at the moment, GIS software is widely available and capable of importing data in various formats. Moreover, the visual quality of recent historical GIS applications is stunning.[61] There is even good software freely available, and, maybe most important, the problem of geographic changes over time has been noted by historians, geographers and information scientists alike.

In a recent introduction to the use of GIS in historical research, Ian Gregory (2002) cites Peuquet who stated that a fully temporal GIS must be able to answer three types of queries:

Changes to a spatial object over time, such as 'has the object moved in the last two years?', 'where was the object two years ago?' or 'how has the object changed over the past five years?'

[61] A good sample of recent historical GIS projects is in (Knowles, 2002).

Changes in the object's spatial distribution over time, such as 'what areas of agricultural land-use in 1/1/1980 had changed to urban by 31/12/1989?', 'did any land-use changes occur in this drainage basin between 1/1/1980 and 31/12/1989?', and 'what was the distribution of commercial land-use 15 years ago?'

Changes in the temporal relationships among multiple geographical phenomena, such as 'which areas experienced a landslide within one week of a major storm event?', 'which areas lying within half a mile of the new bypass have changed from agricultural land use since the bypass was completed?'

Gregory comes to the conclusion that at the moment there is no GIS system that can cope with these three queries. But this does not mean to say that no progress has been made at all in this respect. For instance, regarding the problem of boundary changes between spatial objects over time, a number of solutions did have been devised. Leaving aside the very easy solutions – a new map is drawn every time a boundary change takes place – which is very time-consuming, takes a lot of disk space and does not facilitate easy comparisons of spatial changes over time, there are two ways to solve this problem: by using a date-stamping approach, or by using a space-time composite approach. The data-stamping approach, which for instance has been used by (Gregory and Southall, 2000) and (Boonstra, 1994b), defines time as an attribute to a spatial points and spatial objects. An administrative unit *x* is a spatial point, with a specific starting date and a specific end date as its time attributes, as well as a set of lines, also with specific starting and end dates. When drawing a map of *x* at a specific moment in time *t*, only those lines are selected that have start and end dates on either side of *t*. See Figure 3.21. Boonstra used a similar approach, using polygon attributes instead of line attributes.

The space-time composite approach defines administrative units as a set of smaller polygons that do not change over time. Each polygon has at least one attribute: the administrative unit to which it belongs. If a polygon changes from one administrative unit to another, the attribute data changes as well. These smaller polygons are referred to as the Least Common Geometry (LCG). This can consist of "real" low-level administrative units that are known to be stable over time, as in the Swedish system that uses parishes to create districts, municipalities, and counties (Kristiansson, 2000), but it can also consist of "virtual" polygons that were created as a result of boundary changes. Such a solution has been proposed and tested by (Ott and Swiaczny, 2001) and put to use in HISGIS, the web-based Belgian Interactive Geographic Information System for Historical Statistics.[62] In both cases, a dissolve operation is needed to re-aggregate the polygons in the LCG to form the units in existence at the required time. See Figure 3.22.

[62] More information on the Belgian Historical GIS project at http://www.flwi.ugent.be/hisgis/start_en.htm.

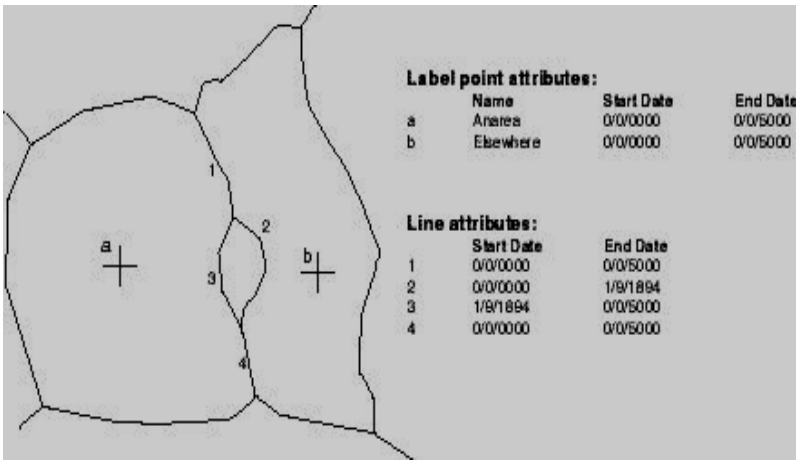


Figure 3.21. Example of the data-stamping approach, showing how a boundary change between Anarea and Elsewhere on 1 September 1894 can be handled. Source: (Gregory and Southall, 2000).

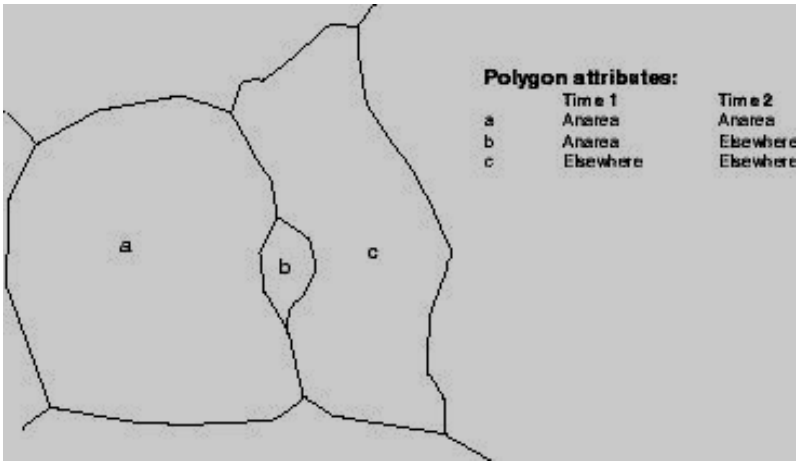


Figure 3.22. Example of the data-space-time composite approach, showing how a boundary change between Anarea and Elsewhere is handled: during Time 1, Anarea is an aggregate of polygons *a* and *b*; during Time 2, Anarea consists of only polygon *a*. Source: (Gregory and Southall, 2000).

3.5.3.6. Conclusions

Internet and low-cost computer storage facilities have triggered much interest in the use of images as a source for historical research. The availability of

hundreds of large collections of digitised images all over the world does not mean that no work needs to be done anymore before serious research can make use of such collections. On the contrary, the possibilities to retrieve images with a sufficient degree of precision and recall are still small.

The inclusion of specific metadata, which deal with the historical meaning and context of the image, in order to cater for the variety in which an image can be interpreted, poses a serious problem for information science to deal with. Trying to develop ways for content-based retrieval of historical images instead of metadata-based retrieval seems to be a spectacular way to circumvent this problem, but it is to be expected that in the near future metadata-based retrieval will generate results that are more interesting for historical research.

Most of the various types of visual data analysis fall within the framework of explorative data analysis, as a means of gaining understanding and insight into the data in a visual way. Although quite a few tools have been developed, there is still much to do in finding ways to present time-varying data visually. Next to that, there is also need for research in which visualisation techniques are developed in combination with tools for exploratory analysis of historical data.

3.6. Final remarks

The structure of this chapter has been based on trends in humanities computing, which has some drawbacks. It leaves us with blind spots, simply because certain techniques or approaches are not explicitly discussed in publications. Multimedia is slowly appearing in humanities research, 'information retrieval' is not labelled as such and mostly addressed implicitly, knowledge in a technical sense (as in 'knowledge engineering' and 'knowledge technology') does not belong to the humanities vocabulary.

The comments made in Section 3.4 on the ambiguity within the ranks of the Association of History and Computing on how to use information technology in historical research, had the intention to make clear that historical information science cannot do without a structural scientific footing, a footing which the AHC has not been able to give. Both content and structure demand conditions that need to be fulfilled to ensure a future for historical information science.

The overview in Section 3.5 was intended to focus on the complexity of computational problems and to indicate handles for applying achievements from, amongst others, computer science, while taking advantage of work that has been done already in the humanities research itself. Text mining and knowledge technology, for example, may open quite new perspectives, but such expertise should be applied on the correct level of abstraction, fitting in with (or at least being aware of) conventional techniques, as record linkage, statistical analysis and source criticism; otherwise new solutions will not be meaningful to the practitioners in the field.

Chapter 4. The Present

4.1. Conclusions from the Past

> "And yet, and yet, while there is much to celebrate about the last decade, the fact remains that the profession is still divided between the small minority of historians who uses computers as tools for analysing historical data and the vast majority who, while they might use a PC for wordprocessing, remain unconvinced of the case that it can become a methodological asset." (Speck, 1994)

These words were written by Speck in 1994. At that time, ten years after the Hull symposium, the position of computing historians had been consolidated; they had become organised and had established their own communication channels. However they had not reached the majority of the profession, which remained resistant to this new branch of methodology.

Moreover, the community was divided in itself. Although scientific discussions would be fruitful and foster intellectual progress in a field, the Association for History and Computing (AHC) had not succeeded in coaxing clear scientific conclusions out of the scholars it addressed. The debate had stopped at the level of being aware of different points of view.

Historical computing had different meanings for different categories of computer using historians. Within the ranks of the AHC a majority supported the idea that information technology was something that simply had to be applied to historical research. They were impressed by the blessings of the 'mighty micro' and convinced of the capabilities of the new technology. For them, historical computing referred mainly to strategies for obtaining historical results with the aid of computers. The precise nature of the procedures to be followed and the limitations of the technology-as-provided were far less important than the results themselves. The methods and techniques deserved only attention as far as they had to be acquired. In this way, this category stayed relatively close to the majority of historians who rejected historical computing as such. From their point of view, historical computing was above all history.

A relatively small number of computer using historians pleaded for historical computing in the sense of a historical information science, and did so with good arguments. Their weakest point, perhaps, was their noble motivation of helping fellow historians to deal with computational problems at an acceptable level of skills without losing focus of their proper historical enquiry. They kept trying to convince their unresponsive colleagues, and unfolded missionary activities in historical projects that could benefit from their expertise. With hindsight, one cannot avoid the conclusion that convincing did not work at all. It must be admitted that in many cases this noble attitude may have been nourished by the understanding that any funding for research could only be obtained through co-operation with regular historical projects – and vice-versa. This

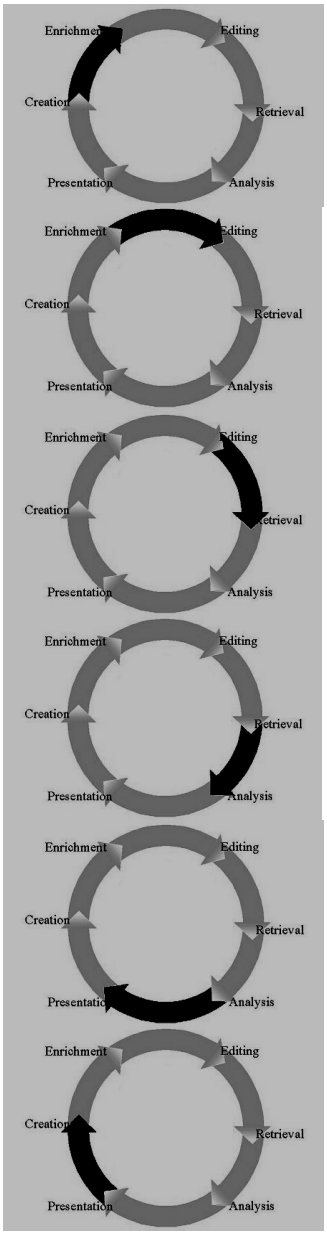
problem was difficult to solve, as we hinted at already in Chapter 3, in the context of the Dutch situation. In addition, historical computing requires a continuous testing of software and techniques under realistic conditions, based on the practice of historical research. Nevertheless, in spite of several promising attempts, this part of the AHC community failed to maintain a clear common focus and to establish broad co-operation.

As a consequence, research into computerised methods and tools for historians to use remained limited. To a large degree, research was either done by outsiders, who did not take part in AHC conferences and did not report to the "history and computing" community, or by individuals who did not find a sounding board for the results they had achieved. In sum, the Association for History and Computing did not live up to the promise of being a platform for researchers in historical information science, nor in disseminating the tools they created to a wider audience of professional historians.

The result of it all is threefold. Firstly, researchers from within the domain of history discussed not all relevant topics and computing - much work was done outside the field. Secondly, a proper research infrastructure for historical information science was not set up. And thirdly, a link between historical information science and general information science failed to become established. These three issues will be discussed in more detail in the next three sections.

4.2. The lost topics

If we look at the themes that have been researched over the last couple of years, it is surprising to see the expanse of issues that has been investigated. If we put all issues mentioned in Chapter 3 along the life cycle of historical information, it is positive to see that all aspects have been touched upon, from creation (creating visual databases for instance), through enrichment and retrieval to analysis and presentation. It is nevertheless evident that within the framework of "history and computing", i.e., the AHC and its journal, some issues have received more attention than others. It is also clear that some issues almost have exclusively been discussed and researched outside the traditional domain of history and computing. Defining such issues as "lost topics", a list can be made of issues that did not get much attention from the domain of history and computing.



	discussed within "history and computing"	*discussed outside "history and computing"*
	creation	
	source oriented data modelling; optical character recognition of old prints and manuscripts	XML modelling; creating visual databases; time-varying historical GIS; textual databases;
	enrichment	
	metadata for historical sources;	XML standards for adding historical metadata; linking source fragments;
	editing	
	record linkage; family reconstitution;	source-critical editions;
	retrieval	
	content based image retrieval;	ontologies for historical research; history & the semantic web; metadata-based image retrieval;
	analysis	
	multiple regression analysis; event history analysis; simulation; statistics on fuzzy data;	exploratory data analysis; visual data analysis; visual text analysis; historical GIS; authorship attribution; content analysis;
	presentation	
	historical GIS;	visual data analysis; visual text analysis; timelines

4.3. A failing infrastructure

As has been stated above, "history and computing" did not succeed in creating an international platform for historical information science. At national levels, attempts to build an infrastructure failed as well. In the Netherlands, for instance, only a very few history departments set up centres for methodological research and development. Consequently, not much thought was given to the formulation of research problems and their solutions, and even less so to the formulation of IT related problems and solutions. In the Netherlands, all national historical research centres like the *Huijgens Institute* and the *N.W. Posthumus Institute* focus exclusively on thematical issues. There is no room for research into methodological issues.

Secondly, there was the short-lived success of History & Computing centres. Twenty years ago, history departments encouraged such centres, in which historians with some knowledge of information technology tried to help their colleagues to cope with the information problems they had. The introduction of easy-to-use Windows-based software resolved a few of the problems specific to history (like for instance the use of relational database systems to help keep track of changes in records over time), causing management officials to think that there was no need for further support any more. University cutbacks have erased History & Computing rapidly and almost completely.

Thirdly, historical research has remained an individually based kind of research. Solutions for historical information problems are therefore always linked to one particular research project, showing a lack of awareness of generalising results.

And finally, when dissemination of IT results in history did take place, it was not among historians, but among the specialists that were working within the History & Computing centres, leaving 'normal' historians unaware of relevant contributions of IT to historical research.

4.4. The failing relation between History and Computing and Information Science

If we look at the list of "lost topics", it becomes obvious that those who adhere to the field of history and computing have not kept themselves abreast of the developments in IT research. For instance, there has hardly been any discussion about the way historical data – with its typical characteristics – could be modelled; no standards have been developed about the way metadata should be added to historical digital data; hardly any research is done into possible new tools for analysis, and so on.

On the one hand, the reason for this is that people working in the field of history and computing do not succeed in disseminating their IT-based solutions to historical problems to the traditional field of historical science. On the other

hand, another reason is that, because "history and computing" did not succeed in creating an international platform for historical information science, there is only a poor relationship between historical information science and information science.

It is not entirely due to the problems encountered within the field of history and computing that there is no relationship between historical information science and information science. It cannot be denied that information scientists rarely give attention to information problems that are typical of historical research. On the one hand, this is because they are not aware that such problems exist, but on the other hand, the idea of solving such problems is not within the realm of problems that are tackled in information science research projects.

Be that as it may, as a consequence, historians and IT specialists have not established a fruitful communication, and therefore cannot at present exchange views, problems and solutions.

Chapter 5. The Future

5.1. Conclusions from the present: A paradox

Up to the present, the field of history and computing has been rather productive in generating publications on the application of computer techniques to historical research. The World Wide Web has changed the way historical sources are published, at least to some extent. Printed editions have met competition from an increasing quantity of digitised sources made available on the Internet, although so far these are only rarely integrated and provided with additional tools as digital archives and digital libraries do.

However, considered as a field of *scientific* study, the position of history and computing is far less favourable. With a few exceptions, methodological and technical problems have not been studied in a generic way, aiming at solutions that could be used in the entire field for a specific class of problems. If this will be realised, the current 'best practice' will pass into a historical information science (Dutch: *historische informatiekunde*), or more precisely, into a specialised branch of information science concerned with the historical domain. As elsewhere in this report 'historical' has to be understood in the broadest sense of the word: regarding 'historical information', not restricted to history as a discipline, but also being the corner-stone of other fields of study in the humanities. At present, it is unlikely that this will happen soon without any change of policies.

The historical community is divided on the question regarding how much emphasis this research direction should get. As long as the phenomenon 'history and computing' is discussed, one paradox has turned up again and again: for historians, the computer will provide usable tools, but somehow these tools

need to be developed. 'The computer' will never help one to write better history automatically, but the analysis of historical problems can be carried out in a more sophisticated and innovative way within the framework of information technology. This is the historian's task, not that of a computer scientist or an information analyst. It is the duty of these 'technicians' to adapt the information technology framework in such a way that historians can do their work: to find new answers to old questions, or even to be able to pose questions that were not thinkable or answerable without the application of information technology. Although this may sound obvious to some of us from our distant point of view in this report, it may be intuitively hard to grasp in the practice of historical studies. Therefore, some elaboration of this paradox may be helpful.

Historians want to write history, and for that purpose they may use a computer to manage and to analyse their data. The computer makes data available, reduces the amount of time spent to searching, and links dispersed chunks of information, allowing questions that require such an extensive quantitative or qualitative analysis that would have otherwise been unfeasible. This leads easily to the conclusion that the activities in historical computing should be directed to practical matters. Along this line of reasoning the value of the solution is demonstrated through the computer-aided historical study itself. Reflection on these activities is considered as useful, but mainly as a form of sharing experiences. Moreover, historians want to stay historians, and do not want to delve into the intricacies of information technology – and they are perfectly right.

The paradox is that enabling the historical community to be practical computer users in such a way, somehow and somewhere a substantial amount of energy has to go to more fundamental methodological and technical research with respect to computing in history. In the early days precisely this has been Thaller's claim, when he developed a special database management system for historical research. It does certainly *not* mean that the every historian has to be turned into a computer expert. There is a parallel with the expertise in editing sources: for example, a historian, who makes use of published medieval charters, can accept very well that some of them are fabrications and deploy this knowledge without being an expert in the technical assessment procedure. However, he should be sufficiently aware of the underlying reasoning and be able to pose adequate questions.

Those historians, who are reluctant to pass the threshold of computer-related methodology and techniques, may be inclined to adapt their questions and research strategies to means readily available. This will result in the use of common commercial software with all its limitations and inconveniences, which, however, will not be necessarily detrimental to the quality of the historical discourse itself. However, even if at first sight standard software seems to suffice, the gap may widen between the solutions thus imposed and the current standards in information technology. But who will tell and who will ac-

knowledge that opportunities are under-utilised? The current tendency to a more narrative history will not easily yield an incentive in this respect.

This leaves us with the thorny questions whether further insistence on methodological and technical research in this field will make any sense, and why the historical information scientist in particular should do so, if the historical community itself is divided and humanities scholars seem to be also happy without more advanced technology. A report on the past, present and future of history and computing inevitably poses the question of scientific progress. Scientific innovation has rarely been based on the consensus of a majority. In the foregoing chapters the development of historical computing has been analysed from a panoramic perspective, highlighting methodological and technical issues in relation to gains in historical knowledge. They open new vistas, showing promising recent developments in computer science and information science and describe successful experiments in larger projects, mostly related to digital heritage and digital libraries or in neighbouring fields, like computer linguistics.

Much of what is blossoming requires further elaboration and has to be translated into more widely applicable and usable tools. Better infrastructure is needed in order to guarantee a transfer of results from the methodological and technical level to the daily practice of historical research. On the contrary, denying these challenges and opportunities will, in the long run, segregate the study of history from the technical capabilities currently being developed in the information society and will turn 'the computer' into an awkward tool with limited use and usability for historians.

5.2. Relevant research lines

In summarising the more detailed information on databases, texts, statistical methods and images from Chapter 3 a few areas of further research stand out (which is not to say that the following list is complete):

1) *Modelling sources and user behaviour; standardisation*. More extensive modelling, of both, the data structure of historical sources and the way sources are used, will greatly aid the interoperability between applications and will make tools more usable. Data modelling applies to the overall structure of sources as well as to data patterns on a micro-level (like references to persons, locations, money, time etc.). In addition, the transformation processes between one data structure and another, as required for specific research purposes, are to be documented through modelling too. All this should be aimed at more uniformity in the data structures and procedures used. The next step will be standardisation, to at least a *de facto* standard.
2) *Supporting editorial processes.* At present most historical and literary text editions use XML. The traditional distinction between typical data-

base data and full text data becomes rather blurred due to novel XML database software. The current generation of XML editors possesses some sophisticated features, but is still rather primitive regarding the editorial process of historical information. Their rigid schema-driven nature may support the encoding of business data which can be completely modelled beforehand very well, but they are less helpful in editing heterogeneous historical material, full of exceptions, which have to be handled on an ad hoc basis. Additionally, more analytical views on the text being edited would be welcome. Finally, a modern edition is no longer bound to the deadlines of printing. Editing can be organised as an ongoing process, realised by means of collaborative software. An edition may become available in instalments, which are reviewed and annotated online by experts all over the world. Organising this process requires new methodological insights, based on additional research.

3) *Discovering structures and patterns.* Apart from critical source editions, which will require a great deal of manual editing, historians will have to cope with lots of raw text. The application of intelligent computer techniques to unstructured texts may be promising, in order to structure texts by creating an elementary form of tagging (semi-)automatically. Next to this is the discovery of patterns and the generation of knowledge through text mining, semantic parsing, content analysis and techniques now summarised under the heading of thematics. A related field of increasing importance is the analysis of images through pattern analysis, possibly in combination with metadata.
4) *Tuning statistical techniques to historical research.* Examples of upcoming statistical techniques more suitable to historical problems have been mentioned in the section on statistical methods above: logistic regression, multilevel regression, event history analysis and ecological inference and new methods for exploratory data analysis with an interactive and visual display of results.
5) *Tuning information retrieval to historical requirements.* Although rarely addressed as such in historical publications information retrieval forms the core of historical information processing. Information retrieval is a well-studied field in computer science; however the complex semantics of historical data and the dimensions of time and space make special demands. Information retrieval is closely connected with authoring, in particular with the addition of metadata. The use of metadata has had a special focus in the research of digital libraries and digital heritage institutions, but it is still far from clear how these mechanisms can be applied in smaller historical research projects.
6) *Multimedia, reconstruction and simulation*. A large and fascinating conglomerate of different technologies is growing around multimedia: geographical information systems (GIS) as applied to historical data,

imaging techniques in the reproduction and analysis of source texts, 3D-reconstructions of historical buildings and locations, providing a special sense of presence and allowing explorations not feasible in 2D-representations. With the exception of historical GIS these technologies are almost virginal fields of study for historians.

7) *Publishing historical discourse*. The new digital counterparts of printed publications tend to adhere to old conventions, often longer than strictly necessary, thus under-utilising the capabilities of the new media. New standards have to be formulated, for example in online journals, for integrating articles with the related historical data and other resources, which might be published together. Museums and libraries are creating more and more exhibitions online, but the majority have difficulties in going beyond the traditional catalogue comprising pictures with a description. Although the Web invites the use of more explorative structures, these are still rare. More effort has to be put in cheaper engineering of high quality historical content.

The realisation of any of these research areas will require organised co-operation with other information-oriented disciplines, like computer science and information science. The latter domain is difficult to define. It varies from university to university, and covers not only the more practical application of information technology in different fields of society, but it is also affiliated with some parts of the social sciences, like cognitive psychology. In the previous chapters we have used the information lifecycle (well known in information science) as an organising principle. Here, we want to summarise these research lines in a slightly different way: as a conceptual framework, comprising several layers, which are closely interlinked (and therefore difficult to represent in a 2D-diagram).

The blocks in the middle represent major topics in computer science and information science. These are also relevant to historical computing, if the typical factors – characteristic of historical information problems (on the right side) – are taken into account: time, space and semantics.[63] 'Content creation' pertains to the scholarly preparation and editing of digitised historical source material. The intelligence layer is a variegated set of intelligent information techniques, aimed at adding structure and classification to, and deriving knowledge from, the historical content. 'Selection' refers to a broad field of information retrieval techniques, which will operate on the already enriched and structured content. Finally, the presentation layer addresses the delivery of information in accordance with the user's level of interest and preferences.

[63] Here, 'semantics' refers to the problem that in historical sources it is often unclear what a source fragment means, see for details Chapter 3.

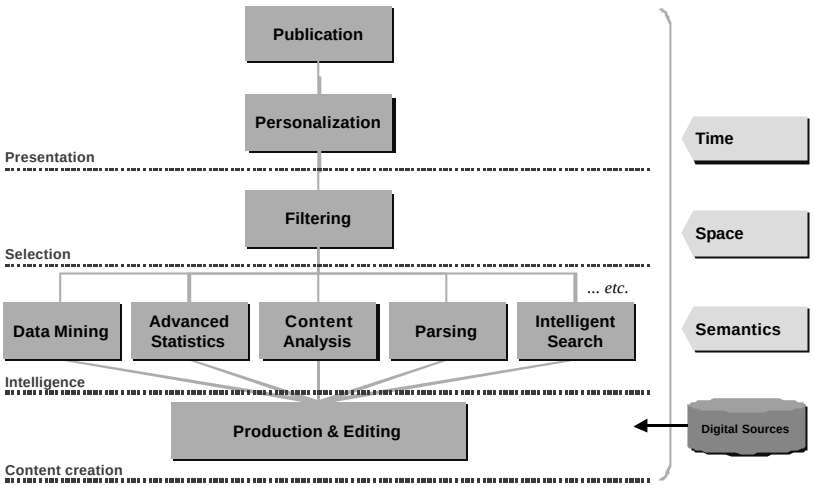


Figure 5.1. Conceptual research framework for a future historical information science.

Some of the research themes mentioned are positioned on a single layer (like the tuning of statistical techniques), while others will mainly belong to a certain layer, but will need also expertise from another (like the support of the editorial process, which is to be primarily classified as content creation, but has aspects of selection and presentation as well).

5.3. A future infrastructure for historical information science: Resolving the paradox

5.3.1. Stakeholder communities

If we want to resolve the aforementioned paradox, then it is necessary to create a situation where this may happen. This pertains directly to mutual relationship between the different disciplines and institutions involved and the infrastructure of scientific collaboration. In the Netherlands the following stakeholder communities could play a vital role in a new information technology offensive in the humanities:

- Cultural heritage institutions, e.g., the Koninklijke Bibliotheek (Royal Dutch Library), the Nationaal Archief (National Archive), Instituut voor Nederlandse Geschiedenis (Institute of Netherlands History), In-

ternationaal Instituut voor Sociale Geschiedenis (International Institute of Social History), major museums, etc.
- Computer-aided projects in humanities, in particular in history, literary studies, archaeology, etc.
- Computer scientists with interest in cultural applications
- (Historical) information scientists

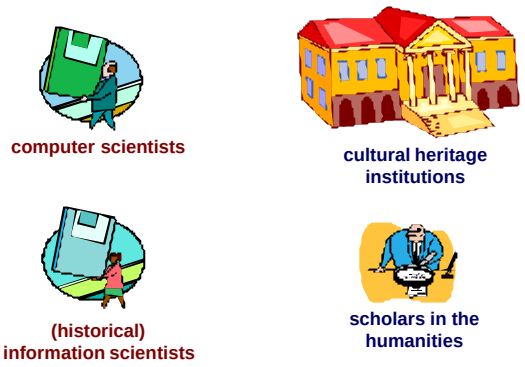


Figure 5.2. Stakeholder communities in humanities computing: relatively is olated and thus under-utilising the field's capacity.

5.3.2. Patterns of co-operation

In the past several smaller and larger projects have existed, in which some of the groups mentioned above have managed to collaborate successfully.

What we would like to see, is a broader span of collaboration (Figure 5.3), which includes all stakeholder communities mentioned above, and leaves room for the participation of those scholars in the humanities working on interesting information problems and wanting to see fundamental solutions applicable beyond their own projects. If at the same time computer scientists and information scientists join forces, the envisaged future will come close to reality and historical information will stand a real chance of succeeding.

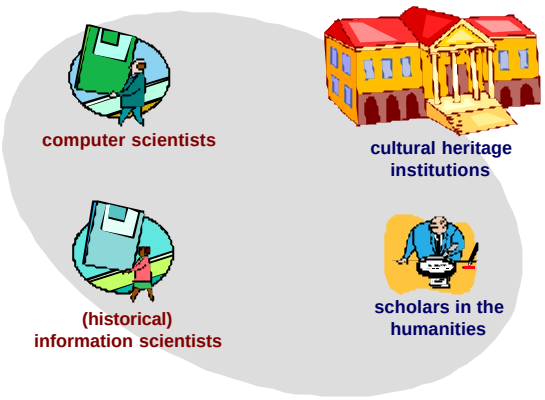


Figure 5.3. Fuller co-operation between the stakeholder communities.

5.3.3. Project models

Collaboration has to be organised. The system of scientific funding and the way research proposals are assessed is highly relevant and may yield a powerful incentive towards a new direction. Furthermore the organisational structure of projects will be a critical success factor. Figure 5.4 shows a frequently used model: technical experts are dropped into a computer-aided project in historical research or in a cultural heritage institution. This model may work very well, as long as 'technicians' restrict themselves to offering services, for example, in the form of system analysis and programming. Problems may arise, when the former group wants to do their own research, using the data and user groups in the hosting project, while the project itself is bound to a timely delivery of content. Methodological and technical research activities seldom keep the same pace as the content-driven activities. Moreover, both parties may disagree about the direction in which the project should evolve. At this stage the project's goal is split. Considering the deplorable state of history and computing as a scientific field, we would not recommend this combination of technical and historical research together within a research project.

A better alternative is shown in Figure 5.5, where two interrelated projects are envisaged: one project, run by computer scientists and (historical) information scientists having specific methodological-technical goals, closely connected with a computer-aided project in history, literature or any cultural heritage institution.

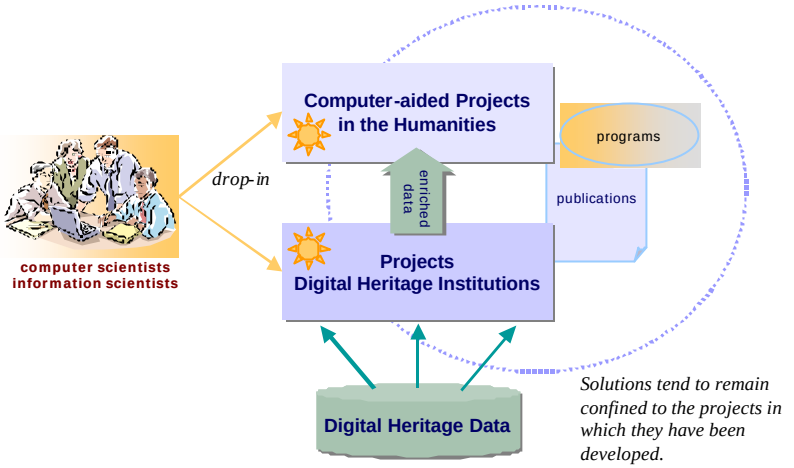


Figure 5.4. A familiar collaboration model, not preferred in this context.

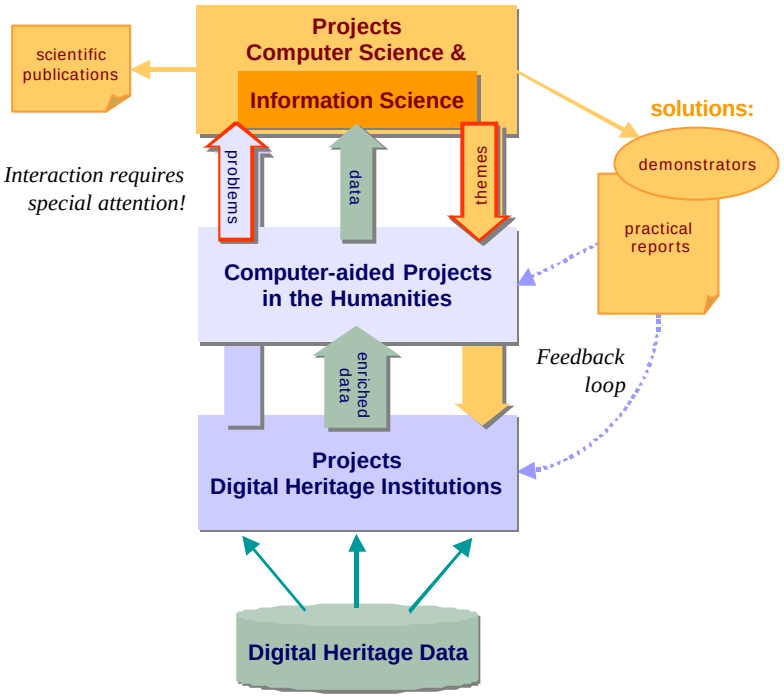


Figure 5.5. Recommended dual-project model for collaboration. This leaves room for parallel research tracks, linked, but without impeding each other.

The success of the second model will depend on a few critical factors. These should be explicitly mentioned.

– *Effective communication*
Good communications between each project is required relating to the research themes and the problems under scrutiny (arrows with a red border). The technical project should propose themes that are of interest for the people involved within that element of the project, but these themes should not be carried forward without a careful taking considerable note of the real problems in the partner project. This process will be necessarily iterative and will require a substantial amount of time. Organising it will be a major management task in itself. An important precondition is a mature state of the historical project: the people there should have explored their problems sufficiently and know already where they are heading. Furthermore, the computer-aided project should be adequately represented in the project team (or steering committee) of the technical project, to guide and to watch over the direction and effectiveness of the problem solving.

– *Scientific publications outside the humanities*
The technical project will have produce regular scientific publications, which may not be of primary interest for scholars in the humanities. They have to do this in order to continue their own scientific activities, or, in the case of historical information science, to establish this activity as a scientific field.

– *Feedback to the humanities*
If this *collaboration* is to mean anything, the technical project will have to make its results available to the partner project in a usable form. The scientific publications mentioned above will not be sufficient. They will likely to be too far away from the daily research practice in the humanities. Therefore, important additional deliverables are demonstrators and technical reports that discuss implementation in detail. A demonstrator in this context should be a working system and not simply a mock-up. However, the right balance will have to be found and the technical project must not slide towards programming on demand. The results may be still beyond the horizon of many less computer-literate scholars in historical and literary studies, but this kind of output should be something to go by for supporting technical staff in arts faculties and cultural heritage institutions. Thus, in a more indirect way, it will support the field as a whole.

5.4. Is that all there is?

This report started with a rather gloomy view on the scientific status of history and computing, and this evaluation has been repeated several times in this overview. On one hand, we have painted the rich and varied landscape of com-

puter applications in history and in the humanities at large and how this field has evolved over the past two decades. On the other hand, the methodological progress was less energetic than it might have been. From the mid 1990s onwards, graphical user interfaces for operating systems and sophisticated 'office' software have provided the illusion that computing could be easy for a large community of historians and could help them to conquer their basic resistance to technology. The Internet has brought scholars and their data together and made distributed, large-scale projects feasible. Digital archives, digital libraries and rich websites of all kinds of cultural heritage institutions became the nodes in a large information web. These focal points have particularly contributed to the more optimistic side of historical computing in the recent past. Large-scale projects have been most fruitful in developing techniques and tools. However, this has often been because the large amount of data to be processed has required automation, and thus justified the development of tools, and the creation of a representative test bed.

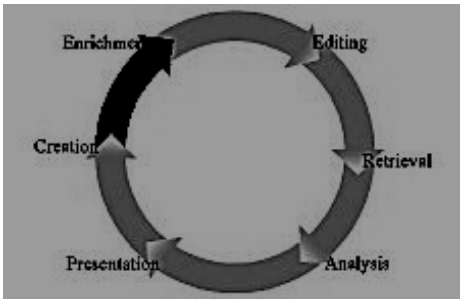
Now we are at the edge of a new development, which may unite those who are interested in information problems (the computer scientists and information scientists) and those who are the treasurers of a wealth of information problems (the scholars in the humanities). As McCarty pointed out, interdisciplinary activity is not a matter of simply importing or exporting ideas and methods, but it is constituted by a unifying perspective at the intersection of two or more fields (McCarty, 2001). Taking two important disciplines in this field as an example, history and computer science, the unifying computing perspective lies in reaching an intermediate level of abstraction with regard to formulating problems and solutions. Historians are inclined to overestimate the uniqueness of their problems, while computer scientists live with the beauty of universal solutions for rather abstract problems. Problems will grow more complex (but also more meaningful) when defined closer to practice. Finding the right balance will require an organisational context where scientists of different denomination meet and work together. We hope that we have outlined such a context convincingly.

One final comment must be added. Any structural change in mentality will require a substantial effort in education. If the envisaged new infrastructure can be realised, a structural co-operation with both undergraduate teaching in universities and with postgraduate research schools must be set up to consolidate this new approach and to connect historical information science with the lifestyle of a new generation which has grown up with sophisticated information technology.

Appendix: Possible themes for historical information research

In Section 5.2, an outline was sketched for a new research program in historical information science. Below a number of themes are presented which fit well within the range of research topics such a research program could address. The themes are arranged according to the life cycle of historical information. It is important to note that they do not comprise a full-grown research program; they need to be thought of as examples, nothing more and nothing less. But they give some idea about the issues that could be handled, as a field of scientific research in which historians cooperate with information scientists, on a level of abstraction which makes it possible to make generalisations and tools that can be disseminated to a large audience of historians.

1. Creation



1.1 Generic modelling of historical cultural sources

Historians have long had access to source commentaries describing the characteristics of certain types of historical sources, like parish registers of baptisms, weddings and funerals, legal deeds and estate inventories, to name just a few. As an extension of this, it must be possible to compile generic data models for sources – in the same way as TEI has designed generic DTDs for literary texts. The term "generic" refers to the data structure shared by different variants of the same main source type, and to a basic functionality which characterises the source type in question. If such generic data models were available to us, data models for specific research exercises would be quicker to produce, and the resulting data collections easier to share.

Because the technologies used for processing texts and those for processing traditional database information have ceased to be so clearly distinguishable since the arrival of XML, it is advisable that these models be defined at a high

conceptual level (with UML class diagrams, for example), not directly related to a particular technology.

There are also more fundamental questions. The practicability of TEI standards is a matter of debate, not only due to their prescriptive character but also because of questions about the nature of text (hierarchical structure versus network structure). Moreover, one can ask how deep and how detailed tagging should be, and what should be done by the editor and what by the user-researcher. Dynamic modelling should also be possible, based upon the researchers' interactions with the text. The good thing about this is that XML is used to express the model, but the model itself is becoming much more flexible, is able to adapt to different interpretations and no longer overburdens the researchers.

QUESTION:
How can historical cultural source material be modelled generically in such a way that physical data models are produced more quickly and uniformly? How can we actually create and publish a series of generic models?

> SUB-QUESTION:
> The modelling of multimedia sources – such as illustrations and music notation – combined with textual material. TEI provides various ways of describing the relationships between text and images in "bimedia" sources, but more specific guidelines are desirable. This issue is also important to historical research programmes like the *visual culture* programme of the Dutch Institute for Netherlands History.

ASSOCIATED THEMES:
4.6

1.2 Ontologies for historical cultural research

Research in the humanities generates new knowledge, which can also be stored in information systems. To do this, however, a system of unambiguous terms – an "ontology"[64] – is needed. There are currently several projects under way which add specific information, in the form of metadata, to existing digitised sources. Much of this work is done at the heritage institutions which hold the original source material.

[64] Ontology: "An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. It implies the hierarchical structuring of knowledge about things by subcategorising them according to their essential (or at least relevant and/or cognitive) qualities. A set of agents that share the same ontology will be able to communicate about a domain of discourse without necessarily operating on a globally shared theory." (Based upon *Hyperdictionary*).

Following on from this, and building upon it, researchers in the humanities should be able to provide their own products with a terminology that makes knowledge accessible through search engines and enables a more effective execution of search queries. The key to designing new ontologies and expanding existing ones is finding domain-specific applications of accepted standards. The actual usability of the design by the target group requires particular attention.

QUESTION:
How can existing techniques for making knowledge accessible, and in particular those related to the design of ontologies, be applied to the practice of historical cultural research?

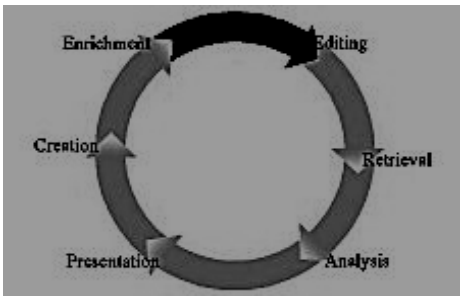
ASSOCIATED THEMES:
2.1, 4.2, 4.7

1.3 Data models for metadata

Research into the usability of RDF and Topic Maps for the storage of metadata. Use of these data models in combination with other – XML and non-XML – formats and models. For example, METS. The interoperability of different systems, such as GIS and thesauri, is associated with this.

QUESTION:
How can metadata be modelled?

2. Enrichment



2.1 The embedding of historical cultural knowledge and insight so as to enrich information systems

The meaning of objects and terms from the past can only be understood in a historical cultural context. Therefore, that context must be included when en-

riching information systems that contain or describe cultural heritage digitally. Research is needed into the question how this can best be done.

If it proves possible to "store" the historical cultural context as metadata, then a further problem soon arises: the addition of metadata can be accelerated considerably if the extraction of terms and the attribution of meanings to metadata is automated. In the case of historical cultural data, this can only be done when the historical cultural context of a source is included in the process of extraction and labelling.

QUESTION:
How can historical data be provided with metadata pertaining to its historical context? If metadata can be provided with metadata pertaining to its historical context, is it also possible to extract and label the metadata automatically?

ASSOCIATED THEMES:
1.2, 5.4, 5.5

2.2 Grammatical help

When developing thesauri or other types of metadata, the meaning of a term can be defined more precisely if its grammatical position is included. So, when adding metadata in order to index information systems, it is advisable to investigate how the addition of such information on the grammatical position of terms can be achieved and what improvements in quality might result from it. This is part of a broader topic: the parsing of a text and the use of the information thus obtained to make that text more useable. For example, there is a particular need for a parser of Middle Dutch and for a diachronous parser that can deal with Dutch from different periods.

QUESTION:
How can the grammatical position of terms be used to refine their meanings, thus resulting in greater precision in their indexing and easier access to the text?

ASSOCIATED THEMES:
5.4

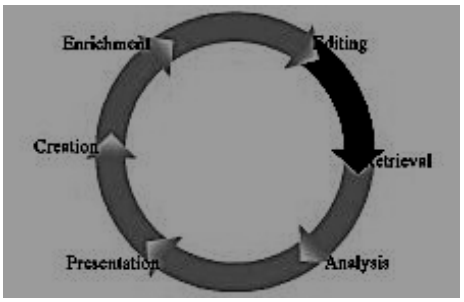
2.3 Changes in the meaning of information over time

For historians, time is a very relevant context. But much of the historical information does not take this into account or does so only implicitly. A trade can shift in relevance or status, a location can change in size or geographical context. How do you include such time-sensitive context, which is not known in advance, in an information system?

QUESTION:
How can changes in the meaning of terms over time be taken into consideration when designing a historical information system?

ASSOCIATED THEMES:
4.2, 4.5

3. Editing



3.1 From historical source to XML text

The current technique for recording textual sources uses XML. Editing an XML text is a fairly technical task, requiring considerable knowledge of the mark-up language when standard XML editors are used. Those who have to do this work because they possess the necessary subject knowledge may well not have the required technical expertise. Research into user-friendly procedures for converting source material to XML form therefore seems desirable.

The existing technology does make it possible to create a modified XML editing environment, but this requires considerable initial technical effort. In addition, it presupposes a more-or-less fixed data model. But historical data tends to be inconsistent and full of exceptions. One possible solution might be found in a modified form of editing, whereby it is easy to change the data model as one works, with the editor itself providing assistance, thanks to its own in-built intelligence.

QUESTION:
Is it possible to design an XML editor tailored to the functional requirements of editing historical cultural sources?

ASSOCIATED THEMES:
5.4, 6.1

3.2 Architecture of digital publication

The formulation of standards for reliable digital publication. What should such a publication look like, which data structure is desirable and which functions should be available to the end user?

QUESTION:

Which requirements should be met by a source or text edition in digital form?

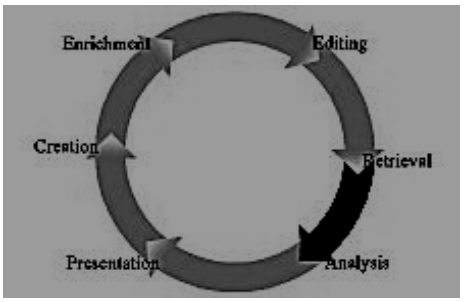
3.3 Publication of sources as an ongoing online process

The prominence of the Internet and the creation of broadband networks make it a realistic proposition to work jointly on the publication of sources at different locations and to regard editing as an ongoing process – with interim results being made available to users immediately. The continual publishing of illustrative material requires particular consideration in this context. In other disciplines, online co-operation has already become commonplace in the form of so-called "collaborations". In particular, the joint online annotation of illustrative material should be investigated further. A related topic is the availability of tools for virtual communities.

QUESTION:

Is joint online work on texts for publication feasible and desirable, and which functionality is required to achieve it?

4. Retrieval



4.1 Historical search strategies

How do historians search for their information? What strategies do they use? Are these strategies based upon traditional handbooks such as *Zoeken en Schri-*

jven (Search and Write), or have new ones been adopted, using the new forms of information gathering offered by the Internet? If the latter were the case, would it be wise to analyse these new search strategies and to evaluate their effectiveness? Moreover, there is also a new and growing group of interested lay people.

QUESTION:
Which search strategies are used by historians, and which of these are most effective? What implications does this have for search systems and their user interfaces?

ASSOCIATED THEMES:
4.3.

4.2. Semantic analysis of search queries

When developing search procedures for the historical cultural domain, changes of meaning and relevance over time must be taken into consideration. A possible way of doing this is to apply a semantic analysis of the search query. This technique translates the original query into a new one, which conveys the user's intention to the underlying information system more accurately. The use of ontologies in search queries is worth considering.

QUESTION:
Is it possible to consider changes of meaning and relevance over time when formulating search queries on history? Could semantic analysis be a solution?

ASSOCIATED THEMES:
1.2, 2.3

4.3 Linguistic search assistance

Simple searches using strings and "wild cards" are often inadequate. More generically applicable solutions for the reduction of search terms to word stems, searches using stems and retrograde searches – that is, based upon word endings – are desirable. Another option to consider is the use of pattern matching. Such solutions are available to some extent in commercial packages. As yet, however, there are no package-independent solutions tailored specifically to historical literature research.

QUESTION:
How can the search process, and textualisation in particular, be improved from a linguistic point of view?

4.4 Text mining, automatic tagging and content analysis

There is far more interesting source material available than can be marked up by hand. This fact forces us to consider the idea of information retrieval from "raw" text – using text mining, for example. Experience with techniques of this kind is mainly confined to modern texts. Very little research has yet been done on their application in historical investigation. In line with this is the question to what extent information can be marked automatically in "raw" texts. The progress with part-of-speech taggers is encouraging. Can the step up to the semantic level be made? In this respect, we should also look at recent research into content analysis and thematics.

QUESTION:
To what extent can information be gleaned from texts which have not been marked up manually?

4.5 Question-answering techniques

When answering factual questions, it is better to produce a factual answer than to simply generate a list of documents in which the answer may – or may not – be found. Question-answering techniques offer this potential. But in the field of historical cultural studies, a factual answer may sometimes be quite correct but only half the story. Is it possible to determine, automatically or otherwise, when a question-answering technique is appropriate and when a more wide-ranging answer is required?

In addition, question-answering techniques will have to be able to include changes in meaning and relevance over time in answers to questions of a historical nature.

QUESTION:
Is it possible to reveal when question-answering techniques should be used to answer search queries, and when they should not? Is it possible to take changes of meaning and relevance over time into consideration when using question-answering techniques?

ASSOCIATED THEMES:
2.3, 4.1.

4.6 Searching multiple information systems

In order to be able to search multiple information systems with different structures and enrichment methods, it is necessary to develop systems that enable interoperability. This is no easy task, as historical sources are retrieved from

information systems in many different ways. Moreover, those systems can contain various types of data: running text, tables, images, audio or any combination thereof. To be able to search the sources effectively, procedures are required which can cope with the variety of retrieval systems used and the many possible combinations of text, images and audio.

QUESTION:
Can procedures be developed for the retrieval of information from multiple information systems? If so, can these be made independent of the type of media upon which those systems are based?

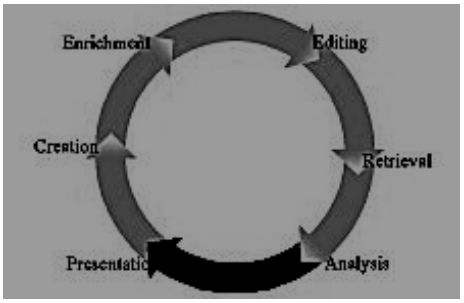
4.7 User profiles during searches

Historical information systems will be used by a wide range of people for a variety of reasons. Based upon the search queries submitted, a system can compile an "ad-hoc" user profile that enables the answers to the questions posed to be tailored to that user's knowledge domain ("adaptive hypertext"). But, in addition to this "ad-hoc" one, it is also possible to have a system – either on its own or in consultation with the user – gradually build up a query-independent user profile. Users can submit this profile at a later time or when visiting other search systems, so that the answers to their questions are tailored to their knowledge.

QUESTION:
Is it possible, based upon the search queries submitted, to compile an ad-hoc user profile which enables those queries to be answered more precisely? And is it possible to compile a permanent, query-independent user profile?

ASSOCIATED THEMES:
1.2.

5. Analysis



5.1 Analysis of numerical data: multi-level statistical techniques

Multi-level techniques make it possible to incorporate variables into a statistical analysis at a higher aggregated level and to produce good estimates of their effects. In this way aggregated data about, for example, a research subject's place of residence – its population, geographical location and economic circumstances – can be combined with individual variables such as gender and year of birth. As of yet, this technique is not used very much in historical study, even though the data required for it is generally available. It is therefore worth investigating the technique's potential in historical research. Such a study should incorporate trials of methods for linking micro-level and aggregate-level data.
Analysis of the opposite type is also possible: research at the aggregate level incorporating information about individual people. Again, such "ecological inference" research is still rare in the study of history.

QUESTION:
What potential is there for the application of multi-level record linkage and multi-level regression analysis in quantitative historical research?

5.2 Analysis of numerical data: event history analysis

Event history analysis is a technique for calculating the chances of a particular event occurring, based upon circumstances and events experienced by the research subjects during their lives. This appears to be a good technique for quantitative research into life histories. But the method of analysis does involve certain assumptions which might be at odds with historical reality. If so, what are the consequences and can they is overcome?

QUESTION:
What potential is there for the application of event history analysis in quantitative historical research?

5.3 Analysis of numerical data: simulation techniques

One of the positive aspects of historical research is that the outcome of the process being studied is already known. This opens up particular opportunities for the evaluation of simulations, and therefore for the way in which this technique can be applied.

QUESTION:
What potential is there for the application of simulation techniques in historical research?

5.4 Analysis of textual data: text-analysis techniques in historical research

Perhaps one of the most promising text-analysis techniques in the study of history is content analysis – providing that the historical context is incorporated into the analysis. But there are also other forms of text analysis which have seldom, if at all, been applied to historical research. Many of these are based upon large written corpora, the grammatical structure of which is already known. With the appearance of historical text files, improved methods of revealing the grammatical structure of texts and the use of XML to record this information as metadata, it is becoming possible to conduct more qualitative textual analysis. This type of research is now attracting many historians. In this respect, so-called "textual statistics" merit particular consideration.

QUESTION:
What potential is there for the analysis of "enriched" texts in historical research?

ASSOCIATED THEMES:
2.1, 3.1.

5.5 Image analysis: content-based image-research techniques

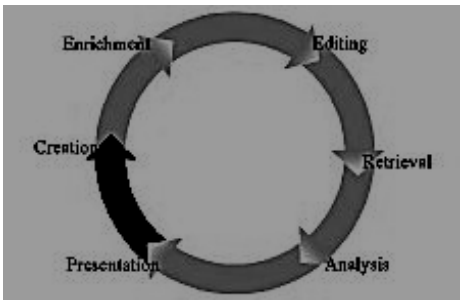
Image analysis uses techniques designed to extract content-based "meaning" from one or more images by means of digital analysis. For example, "this image contains a church/a tree". But can this technique also be applied to historical images – to state, for example, "this image contains William of Orange/Muiderslot Castle".

The technique just described uses only whole or partial images for its search and analytical procedures. This application seems rather limited, though. On the other hand, its combination with textual information opens up many additional possibilities – particularly in the analysis of historical images. So there is every reason to investigate how such contextual information could be used in analysing images.

QUESTION:
Is it possible to use content-based image-research techniques for the analysis of historical images? What potential is there for image analysis using contextual information?

ASSOCIATED THEMES:
2.1.

6. Presentation



6.1 Dynamic generation and presentation of historical information

Texts structured using XML can be broken down into sections. These so-called "components" can then be presented in different forms and context, in accordance with the user's wishes. This opens up a range of opportunities. The system can imperceptibly record the user's choices and modify the presentation accordingly ("adaptive hypertext"). Alternatively, the user can specify what form the presentation takes ("dynamic content"). If the information is to be presented on the web, the location of this process can also vary: either the provider's server or the user's PC.

In order to reach this point, thorough research is required into the structure of the textual material to be used – perhaps involving such things as genre theories or Rhetorical Structure Theory– so as to create models from which the computer can synthesise presentations.

QUESTION:
How can presentations tailored to the needs of a specific audience, or to a large extent configurable by the users themselves, be generated from well-structured historical cultural material?

SUB-QUESTION
The above can be expanded to include the generation of "virtual exhibitions" of multimedia material. The rare examples of successful web projects show that such an exhibition can be much more than "pictures with text". Exploration, animation and elements of play can be incorporated into the "story", producing a far more intensive experience. Although staging a virtual exhibition is an artistic activity, which has little direct relationship with algorithmisation, the computerisation of its components certainly seems a subject worthy of research.

ASSOCIATED THEMES:
3.1.

6.2 Visualisation of historical research subjects and results

By applying information technology, it is possible to recreate a visual representation of historical data, which has come down to us in written form but actually refers to a past three-dimensional reality.

- Historic buildings can be reconstructed on screen using 3D techniques, giving the user a much more intensive experience than could be obtained from studying the sources.
- Animations can be used to enhance the study of works of art (on-screen rotation, panoramic views, exploded views et cetera).
- Co-ordination of historical data in time and space:
- geographical information systems (GISs) have been in use for the historical study of cities, towns and regions for some time now;
- Timelines – which never really came into their own in printed historical atlases because of the medium's limitations – can be shown on screen in a rich variety of forms.
- Visual representation of textual structures.
- Different "views" of the same text, each with its own structure and including or omitting certain information in order to highlight certain aspects, visualise the frequency of textual properties or show the structure of a manuscript.

This theme encroaches upon a relatively new field, much of which is yet to be mapped out. Not only is there still a lot of interesting detailed research to be carried out, but a clear overview is also lacking.

QUESTION:
Which main themes in visualising historical material, in the broad sense of the term, are still to be revealed? Can historical cultural functionality be formulated generically? With which aspects of information technology do those themes correspond? How efficient is the available software?

7. Central themes

7.1 Digital durability of research products

Digital durability is currently the subject of considerable attention, particularly with respect to the archiving of computerised administrations. Another aspect is the durability of digital research products such as databases and digital sources with embedded scripts. The proper preservation of these for future use must be guaranteed, which requires effective archiving of digital historical data.

7.2 Usability of software tools

As a rule, researchers working in the humanities use off-the-shelf software products. There are only a very few specialist programmes developed for this field of study. An inventory needs to be compiled covering both categories, particularly with regard to their range of applications. A website on historical software would be particularly useful! Research is also needed into the question how useful particular tools are for the tasks they are actually carrying out. By considering usability more systematically, the need for specifically-developed software tools will become clearer.

7.3 Modelling of historical cultural information systems

The methods and techniques used to design information systems are still largely based upon situations in the commercial world. There, data structures and desired functionality are usually determined within the organisation itself. The design follows the structure of the work process – which we can reasonably assume is fairly stable, at least for a certain period.

Information systems for historical cultural research, on the other hand, are characterised by the fact that whilst – apart from exceptions caused by the heterogeneous nature of the source material – the overall structure of the material is clear, functional requirements tend to evolve relatively rapidly. After all, as the understanding of the subject increases it prompts new questions and processes. The tool must be able to evolve at the same pace, a factor which needs to built into the design and incorporated into the project plan.

It would be useful to formulate a framework of criteria, using which IT-related historical cultural research projects can be described and evaluated.

SUB-QUESTION:
Content management in historical cultural research: Content management systems are used by organisations to store, administer, process and publish all kinds of data. A wide of range of such software is available commercially. What all the packages have in common is that they support the information lifecycle. Once we succeed in modelling the lifecycle of historical cultural research – at a certain level of abstraction – then it will become

possible to better harmonise and integrate different software tools. A content management system for this academic sector should be virtual in nature, perhaps consisting of a series of separate but co-ordinated and standards-based programs.

References

(1993). *Information technology in humanities scholarship: British achievements, prospects and barriers*. London: British Library and British Academy.

E. Aarseth (1998). 'From Humanities Computing to Humanistic Informatics: Creating a Field of Our Own', *Paper presented at The Future of the Humanities in the Digital Age, Bergen, Norway*.

K. Abram (2002). 'Electronic Textuality. A bibliographic essay'. Derived from the World Wide Web: http://www.mantex.co.uk/ou/resource/elec_txt.htm.

P. Adman (1997). 'Record Linkage Theory and Practice: A Matter of Confidence', *History and Computing* 9 (1-3): 150-155.

P. Adman, S.W. Baskerville and K.F. Beedham (1992). 'Computer-Assisted Record Linkage: or How Best to Optimize Links Without Generating Errors', *History and Computing* 4 (1): 2-15.

R.C. Allen (1998). 'Capital accumulation, the Soft Budget Constraint and Soviet Industrialisation', *European Review of Economic History* 2 (1): 1-24.

R.C. Allen and I. Keay (2001). 'The first Great Whale Extinction: the End of the Bowhead Whale in the Eastern Arctic', *Explorations in Economic History* 38 (4): 448-477.

P.D. Allison (1984). *Event History Analysis. Regression for Longitudinal Event Data*. Beverly Hills / New York: Sage.

G. Alter (1998). 'L'Event History Analysis en Démographie Historique. Difficultés et Perspectives', *Annales de Demographie Historique* (2): 25-50.

G. Alter and M.P. Gutmann (1999). 'Casting Spells Database Concepts for Event-History Analysis', *Historical Methods* 32 (4): 165-176.

A. Andreev, L. Borodkin and M. Levandovskii (1997). 'Using Methods of Non-Linear Dynamics in Historical Social Research: Application of Chaos Theory in the Analysis of the Worker's Movement in Pre Revolutionary Russia', *Historical Social Research* 22 (3-4): 64-83.

M. Artzrouni and J. Komlos (1996). 'The Formation of the European State System a Spatial 'Predatory' Model', *Historical Methods* 29 (3): 126-134.

S. Aumann, H.-H. Ebeling, H.-R. Fricke, et al. (1999). 'From Digital Archive to Digital Edition', *Historical Social Research* 24 (1): 101-144.

R.H. Baayen, H. van Halteren and F.J. Tweedie (1996). 'Outside the Cave of Shadows. Using Syntactic Annotation to Enhance Authorship Attribution', *Literary and Linguistic Computing* 11 (3): 121-131.

J. Bacher (1989). 'Einführung in die Clusteranalyse mit SPSS-X für Historiker und Sozialwissenschaftler', *Historical Social Research* 14 (2): 6-167.
J.T. Barry, S. Walby and B. Francis (1990). 'Graphical Exploration of Work History Data', *Quad. Statist. Mat. Appl. Sci. Econ Sociali.* 12: 65-74.
T. Bengtsson (1999). 'The Vulnerable child. Economic Insecurity and Child Mortality in Pre-industrial Sweden: a Case Study of Västanfors, 1757-1850', *European Journal of Population/Revue Européenne de Démographie* 15 (2): 117-151.
T. Bengtsson and G. Broström (1997). 'Distinguishing Time Series Models by Impulse Response: a Case Study of Mortality and Population Economy', *Historical Methods* 30 (4): 165-171.
T. Bengtsson and M. Dribbe (2002). 'Fertility Response to Economic Stress. Deliberate control or Reduced Fecundability?' *Lund Papers in Economic History* 78.
T. Bengtsson and M. Lindström (2003). 'Airborne Infectious Diseases during Infancy and Mortality in Later Life in Southern Sweden, 1766-1894', *International Journal of Epidemiology* 32: 286-294.
G. Bloothooft (1994). 'Corpus-based Name Standardization', *History and Computing* 6 (3): 153-167.
G. Bloothooft (1995). 'Multi-Source Family Reconstruction', *History and Computing* 7 (2): 90-103.
G. Bloothooft (1998). 'Assessment of Systems for Nominal Retrieval and Historical Record Linkage', *Computers and the Humanities* 31: 39-56.
H.-P. Blossfeld and G. Rohwer (2002). *Techniques of Event History Modelling . New Approaches to Causal Analysis*. Mahwah, New Jersey: Erlbaum.
O. Boonstra (1994a). 'Automatisering en het Kadaster. Het Gebruik van de Computer bij Historisch-kadastraal Onderzoek', *Cahier VGI* 8: 114-123.
O. Boonstra and M. Panhuysen (1999). 'From Source-oriented Databases to Event-history Data files: a Twelve-step Action Plan for the Analysis of Individual and Household Histories', *History and Computing* 10 (1-3): 1-9.
O.W.A. Boonstra (1990). 'Supply-side Historical Information Systems. The Use of Historical Databases in a Public Record Office', *Historical Social Research* 15: 66-71.
O.W.A. Boonstra (1994b). 'Mapping the Netherlands, 1830-1994: The Use of NLKAART', in: M. Goerke, *Coordinates for Historical Maps*. St. Katharinen: Halbgraue Reihe, 156-161.
O.W.A. Boonstra (2001). 'Breukvlakken in de Eenwording van Nederland', in: J.G.S.J.v. Maarseveen and P.K. Doorn, *Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900*. Amsterdam: IISG, 277-298.
O.W.A. Boonstra, L. Breure and P.K. Doorn (1990). *Historische Informatiekunde*. Hilversum: Verloren.
O.W.A. Boonstra, P.K. Doorn and F.M.M. Hendrickx (1990). *Voortgezette Statistiek voor Historici*. Muiderberg: Coutinho.

K. Börner and C. Chen (2002). 'Visual Interfaces to Digital Libraries. Lecture Notes in Computer Science.' Heidelberg: Springer-Verlag.
L. Borodkin (1996). 'History and Computing in the USSR/Russia: Retrospection, State of Art, Perspectives'. Derived from the World Wide Web: http://www.ab.ru/~kleio/aik/aik.html.
L. Borodkin and M. Svishchev (1992). 'El Sector Privado de la Economia Sovietica en Los Anos Veinte', *Revista de Historia Económica* 10 (2): 241-262.
B. Bos and G. Welling (1995). 'The Significance of User-Interfaces for Historical Software', in: G. Jaritz, I.H. Kropac, and P. Teibenbacher, *The Art of communication. Proceedings of the Eight International Conference of the Association for History and Computing, Graz, Austria, August 24-27, 1993*. Graz: Akademische Druck- und Verlaganstalt, 223-236.
J. Bradley (1994). 'Relational Database Design and the Reconstruction of the British Medical Profession: Constraints and Strategies', *History and Computing* 6 (2): 71-84.
L. Breure (1992). 'Tools for the Tower of Babel: Some Reflections on Historical Software Engineering', in: *Eden or Babylon? On Future Software for Highly Structured Historical Sources*. St. Katharinen: Max-Planck-Insitut für Geschichte, Göttingen, 23-36.
L. Breure (1994a). 'How To Live With XBase: the Socrates Approach', in: F. Bocchi and P. Denley, *Storia & Multimedia. Proceedings of the Seventh International Congress Association for History & Computing*. Bologna: Grafis Edizioni, 477-484.
L. Breure (1994b). 'SOCRATES: Tools for Database Design and Management', in: H.J. Marker and K. Pagh, *Yesterday. Proceedings from the 6th international conference Association of History and Computing, Odense 1991*. Odense: Odense University Press, 140-148.
L. Breure (1995a). 'Altis. A Model-based Approach to Historical Data-entry', *Cahier VGI* 9: 178-188.
L. Breure (1995b). 'Interactive Data Entry: Problems, Models, Solutions', *History and Computing* 7 (1): 30-49.
L. Breure (1999). 'In Search of Mental Structures: A Methodological Evaluation of Computerized Text Analysis of Late Medieval Religious Biographies', *History and Computing* 11 (1-2): 61-78.
M. Brod (1998). 'Computer Simulation of Marriage Seasonality', *History and Computing* 10 (1-3): 10-16.
A.S. Bryk and S.W. Raudenbush (1992). *Hierachical Linear Models. Applications and Data Analysis Methods*. Newbury Park: Sage.
P. Burke (2001). *Eyewitnessing: The Uses of Images as Historical Evidence*. Ithaca: Cornell University Press.

L. Burnard (1987). 'Primary to Secondary: Using the Computer as a Tool for Textual Analysis in Historical Research', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 228-233.
L. Burnard (1989). 'Relational Theory, SQL and Historical Practice', in: C. Harvey, *History and Computing II*. Manchester, New York: Manchester University Press, 63-71.
L. Burnard (1990). 'The Historian and the Database', in: E. Mawdsley, N. Morgan, L. Richmond, and R. Trainor, *History and Computing III. Historians, Computers and Data. Applications in Research and Teaching*. Manchester, New York: Manchester University Press, 3-7.
J. Burrows (2003). 'Questions of Authorship: Attribution and Beyond', *Computer and the Humanities* 37: 5-32.
J.F. Burrows (1987). *Computation into Criticism: A Study of Jane Austen and an Experiment in Method*. Oxford: Clarendon Press.
J. Burt and T.B. James (1996). 'Source-Oriented Data Processing. The Triumph of the Micro over the Macro?' *History and Computing* 8 (3): 160-168.
J.W. Buzydlowski, H.D. White and X. Lin (2002). 'Term Co-occurrence Analysis as an Interface for Digital Libraries', in: C.C. K. Börner, *Visual Interfaces to Digital Libraries*. Heidelberg: Springer Verlag, 133-144.
C. Campbell and J. Lee (2001). ' Free and unfree Labor in Qing China: Emigration and Escpae among the Bannermen of Northeast China, 1789-1909', *History of the Family* 6 (4): 455-476.
C. Campbell and J.Z. Lee (1996). 'A Death in the Family: Household Structure and Mortality in Rural Liaoning: Life-event and Time-series Analysis', *History of the Family* 1 (3): 297-328.
K.D. Cartwright (2000). 'Shotgun Weddings and the Meaning of Marriage in Russia: an Event History Analysis', *History of the Family* 5 (1): 1-22.
C. Chen and K. Bömer. (2002). 'Top Ten Problems in Visual Interfaces to Digital Libraries', in: K. Börner and C. Chen, *Visual Interfaces to Digital Libraries*. Heidelberg: Springer Verlag, 226-231.
T. Coppock (1999). 'Information Technology and Scholarship: Applications in the Humanities and Social Sciences.' Oxford: Oxford University Press for the British Academy.
L. Corti (1984a). 'Automatic processing of art history data and documents'. Pisa: Scuola Normale Superiore.
L. Corti (1984b). 'Census: Computerization in the history of art.' Los Angeles: The J. Paul Getty Trust.
L. Corti and M. Schmitt (1984). 'Automatic processing of art history data and documents. Second International Conference - proceedings.' Pisa: Scuola Normale Superiore.
P. Cowley and J. Garry (1998). 'The British Conservative Party and Europe: the Choosing of John Major', *British Journal of Political Science* 28 (3): 473-499.

G. Crane (2000). 'Designing Documents to Enhance the Performance of Digital Libraries. Time, Space, People and a Digital Library on London', *D-Lib Magazine* 6 (7/8).

G. Crane, D.A. Smith and C.E. Wulfman (2001). 'Building a Hypertextual Digital Library in the Humanities: a Case Study on London', *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*: 426-434.

C. Daassi, M. Dumas, M.-C. Fauvet, et al. (2000). 'Visual Exploration of Temporal Object Databases.' Presented at BDA 2000, Blois, France.

O. Darné and C. Diebolt (2000). 'Explorations in Monetary Cliometrics. The Reichsbank: 1876-1920', *Historical Social Research* 25 (3-4): 23-35.

H.R. Davies (1992). 'Automated Record Linkage of Census Enumerators' Books an Registration Data: Obstacles, Challenges and Solutions', *History and Computing* 4 (1): 16-26.

M. Debuisson (2001). 'The Decline of Infant Mortality in the Belgian Districts at the Turn of the 20th Century', *Belgisch Tijdschrift voor Nieuwste Geschiedenis* 31 (3-4): 497-527.

H.E. Delger (2003). *Nuptiality and Fertility: an Investigation into Local Variations in Demographic Behaviour in Rural Netherlands about 1800*. Hilversum: Verloren.

F. Deng (1997). 'Information Gaps and Unintended Outcomes of Social Movements: The 1989 Chinese Student Movement', *American Journal of Sociology* 102 (4): 1085-1112.

P. Denley (1994a). 'Models, Sources and Users: Historical Database Design in the 1990s', *History and Computing* 6 (1): 33-43.

P. Denley (1994b). 'Source-Oriented Prosopography: Kleio and the Creation of a Data Bank of Italian Renaissance University Teachers and Students', in: F. Bocchi and P. Denley, *Storia & Multimedia. Proceedings of the Seventh International Congress Association for History & Computing*. Bologna: Grafis Edizioni, 150-160.

K. Depuydt and T. Dutilh-Ruitenberg (2002). 'TEI encoding for the Integrated Language Database of 8th to 21st-Century Dutch', in: C. Povlsen, *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002. Copenhagen, Denmark, August 13-17, 2002*. 683-688.

R. Derosas (1999). 'Residential Mobility in Venice, 1850-1869', *Annales de Démographie Historique* (1): 35-61.

C. Diebolt and V. Guiraud (2000). 'Long Memory Time Series and Fractional Integration. A Cliometric Contribution to French and German Economic and Social History', *Historical Social Research* 25 (3-4): 4-22.

C. Diebolt and J. Litago (1997). 'Education and Economic Growth in Germany before the Second World War: an Econometric Analysis of Dynamic Relations', *Historical Social Research* 22 (2): 132-149.

A. Diekmann and H. Engelhardt (1999). 'The Social Inheritance of Divorce: Effects of Parent's Family Type in Postwar Germany', *American Sociological Review* 64 (6): 783-793.

J.d. Does and J. Voort van der Kleij (2002). 'Tagging the Dutch PAROLE Corpus', in: M. Theune, *Computational Linguistics in the Netherlands 2001. Selected Papers from the Twelfth CLIN Meeting*. Amsterdam, New York: Rodopi, 62-76.

P. Doorn (2000). 'The Old and the Beautiful. A Soap Opera about Misunderstanding between Historians and Models', in: L. Borodkin and P. Doorn, *Data Modelling Modelling History. Proceedings of the XI International Conference of the Association for History and Computing, Moscow, August 1996*. Moscow University Press, 2-29.

P.K. Doorn and J.T. Lindblad (1990). 'Computertoepassingen in de Economische Geschiedenis, in het bijzonder bij Tijdreeksanalyse', *Tijdschrift voor Geschiedenis* 103 (2): 326-341.

S. Drobnic, H.-P. Blossfeld and G. Rohwer (1999). 'Dynamics of Women's Employment Patterns over the Family Life Course: a Comparison of the United States and Germany', *Journal of Marriage and the Family* 61 (1): 133-146.

T. Dutilh and T. Kruyt (2002). 'Implementation and Evaluation of PAROLE PoS in a National Context', in: C.P.S. Araujo, *Proceedings of the third International Conference on Language Resources and Evaluation, ELRA*. Paris, 1615-1621.

M.J. Egger and J.D. Willigan (1984). 'An Event-history Analysis of Demographic Change in Renaissance Florence', *American Statistical Association, 1984 proceedings of the Social Statistics Section*: 615-620.

J. Everett (1995). 'Kleio 5.1.1: A source-oriented data processing system for historical documents. Technical review', *Computer and the Humanities* 29: 307-316.

R.W. Fogel and S.L. Engerman (1974). *Time on the Cross*. Boston; Toronto: Little, Brown and Company.

C. Folini (2000). 'How to bring Barzabal Facin on the screen? A student in search of suitable database architecture', *History and Computing* 12 (2): 203-214.

I. Foster and C. Kesselman (1999). *The Grid: Blueprint for a New Computing Infrastructure*. Los Angeles: Morgan Kaufmann.

B. Francis and J. Pritchard (1998). 'Visualisation of Historical Events using Lexis Pencils'. Derived from the World Wide Web: http://www.agocg.ac.uk/reports/visual/casestud/francis/francis.pdf.

F.S. Frey and J.M. Reilly (1999). *Digital Imaging for Photographic Collections - Foundations for Technical Standards*. Rochester: Image Permanence Institute, Rochester Institute of Technology.

J.B. Friedman (1992). 'Cluster Analysis and the Manuscript Chronology of William du Stiphel, a Fourteenth-Century Scribe at Durham', *History and Computing* 4 (2): 75-97.
A.H. Galt (1986). 'Social Class in a Mid-Eighteenth-Century Apulian Town: Indications from the Castato Onciario', *Ethnohistory* 33 (4): 419-447.
N. Gershon and S.G. Eick (1997). 'Guest Editors' Introduction to Special Issue on Information Visualization', *IEEE Computer Graphics and Applications* 17 (4): 29-31.
T. Gevers and A.W.M. Smeulders (2004). 'Content-based Image Retrieval: An Overview Survey on content-based image retrieval', in: G. Medioni and S.B. Kang, *Emerging Topics in Computer Vision*. New York: Prentice Hall.
A. Gilmour-Bryson (1987). 'Computers and Medieval Historical Texts', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 3-9.
M. Goerke (1994). *Coordinates for Historical Maps*. St. Katharinen: Halbgraue Reihe.
P. González (1998). *Computerization of the Archivo General de Indias: Strategies and Results*: Council on Library and Information Resources [Also full-text availabe in HTML].
L. Goodman (1959). 'Some Alternatives to Ecological Correlation', *American Journal of Sociology* 64: 610-625.
R.C. Graul and W. Sadée. 'Evolutionary Relationships Among G Protein-Coupled Receptors Using a Clustered Database Approach'. Derived from the World Wide Web: http://itsa.ucsf.edu/~gram/home/gpcr.
D. Greasley and L. Oxley (1998). 'Comparing British and Americal Economic and Industrial Performance 1860-1993: a Time Series Perspective', *Explorations in Economic History* 35 (2): 171-195.
M. Greenhalgh (1987). 'Databases for Art Historians: Problems and Possibilities', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 156-167.
D. Greenstein (1997). 'Bringing Bacon Home: The Divergent Progress of Computer-Aided Historical Research in Europe and the United States', *Computers and the Humanities* 30: 351-364.
D. Greenstein and L. Burnard (1995). 'Speaking with One Voice: Encoding Standards and the Prospect for an Integrated Approach to Computing in History', *Computers and the Humanities* 29: 137-148.
D.I. Greenstein (1989). 'A Source-Oriented Approach to History and Cpmputing: The Relational Database', *Historical Social Research* 14 (51): 9-16.
I.N. Gregory and H.R. Southall (2000). 'Spatial Frameworks for Historical Censuses - the Great Britain Historical GIS', in: P.K. Hall, R. McCaa, and G. Thorvaldsen, *Handbook of Historical Microdata for Population Research*. Minneapolis: Minnesota Population Center, 319-333.

G.C. Grinstein and M.O. Ward (2002). 'Introduction to Data Visualization', in: U. Fayyad, G.C. Grinstein, and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco: Morgan Kaufmann, 21-46.

E. Gröller (2001). 'Insight into Data through Visualization', in: P. Mutzel, M. Jünger, and S. Leipert, *GD 2001*. Heidelberg: Springer-Verlag, 352-366.

F. Guérain-Pace and X. Lesage (2001). 'Le Systeme Urbain Français. Les Mesures de l'Inégalité de Distributions de Type Partetien', *Histoire & Mesure* 16 (102): 157-183.

M.P. Gutmann and G. Alter (1993). 'Family Reconstitution as Event-History Analysis', in: D. Reher and R. Schofield, *Old and New Methods in Historical Demography*. Oxford: Clarendon, 159-177.

M.P. Guttman (2002). 'Preface', in: A.K. Knowles, *Past Times, Past Place. GIS for history*. Redlands, CA. 2002: ESRI Press,

D. Haks (1999). 'Two Examples of the Impact of Computer Technology om Historical Editing: The Correspondence of William of Orange 1533-1584 and the Resolutions of the States general 1626-1651', *Journal of the Association for History and Computing* 2 (3): pages n.a.

P. Hartland and C. Harvey (1989). 'Information Engineering and Historical Databases', in: P. Denley, S. Fogelvik, and C. Harvey, *History and Computing II*. Manchester, New York: Manchester University Press, 44-62.

R. Hartmann. 'Prometheus. Das Verteilte Digitale Bildarchiv für Forschung und Lehre'. Derived from the World Wide Web: http://www.prometheus-bildarchiv.de.

C. Harvey (1990). 'The Nature and Future of Historical Computing', in: E. Mawdsley, N. Morgan, L. Richmond, and R. Trainor, *History and Computing III. Historians, Computers and Data. Applications in Research and Teaching*. Manchester, New York: Manchester University Press, 205-211.

C. Harvey and E. Green (1994). 'Record Linkage Algorithms: Efficiency, Selection and Relative Confidence', *History and Computing* 6 (3): 143-152.

C. Harvey, E.M. Green and P.J. Corfield (1996). 'Record Linkage Theory and Practice: an Experiment in the Application of Multiple Pass Linkage Algorithms', *History and Computing* 8 (2): 78-89.

C. Harvey and J. Press (1992). 'Relational Data Analysis: Value, Concepts and Methods', *History and Computing* 4 (2): 98-109.

C. Harvey and J. Press (1993). 'Structured Query Language and Historical Computing', *History and Computing* 5 (3): 154-168.

C. Harvey and J. Press (1996). *Databases in Historical Research*. Wiltshire: Antony Rowe.

A. Hayami and S. Kurosu, Regional Diversity in Demographic and Family Patterns in Preindustrial Japan. Journal of Japanese Studies 2001 27(2): 295-321. (2001). 'Regional Diversity in Demographic and Family Patterns in Preindustrial Japan', *Journal of Japanese Studies* 27 (2): 295-321.

E.A. Henderson (2000). 'When States Implode: the Correlates of Africa's Civil Wars, 1950-92', *Studies in Comparative International Development* 35 (2): 28-47.
C.F. Hermann and M.G. Hermann (1967). 'An Attempt to Simulate the Outbreak of World War I', *American Political Science Review* 61 (2): 400-416.
T. Hershberg (1981). 'The Philadelphia History Project.' Philadelphia, NY.
E. Higgs (1998). *History and Electronic Artefacts*. Oxford.
S. Hockey (1999). 'Is There a Computer in this Class?' Derived from the World Wide Web: http://www.iath.virginia.edu/hcs/hockey.html.
A. Hodgkin (1987). 'History and Computing: Implications for Publishing', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 256-261.
D.I. Holmes and R.S. Forsyth (1995). 'The Federalist Revisited: New Directions in Authorship Attribution', *Literary and Linguistic Computing* 10 (2): 111-127.
D.I. Holmes, L.J. Gordon and C. Wilson (1999). 'A Widow and her Soldier: A Stylometric Analysis of the 'Picket Letters'', *History and Computing* 11 (3): 159-179.
J.J. Hox (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ [etc.]: Lawrence Erlbaum Associates.
E. Hyvönen, S. Saarela and K. Viljanen (2003). 'Intelligent Image Retrieval and Browsing Using Semantic Web Techniques - a Case Study.' Presented at SEPIA Conference 2003, Helsinki.
N.M. Ide (1995). 'The TEI: History, Goals, and Future', *Computers and the Humanities* 29: 5-15.
L. Isaac, L. Christiansen, J. Miller, et al. (1998). 'Temporally Recursive Regression and Social Historical Inquiry: an Example of Cross-movement Militancy Spillover', *International Review of Social History* 43 (Supplement 6): 9-32.
A. Janssens (1989). 'Een 'Direct-entry Methodology' voor Negentiende Eeuwse Bevolkingsregisters', *Cahiers voor Geschiedenis en Informatica* 3: 19-41.
M. Jensen (2003). 'Visualizing Complex Semantic Timelines'. Derived from the World Wide Web: http://newsblip.com/tr/.
M. Katzen (1990). 'Scholarship and Technology in the Humanities.' London: Bowker Saur.
K.S.B. Keats-Rohan (1999). 'Historical Text Archives and Prosopography: the COEL Database system', *History and Computing* 10 (1-3): 57-72.
M. Keiding (1990). 'Statistical Inference in the Lexis Diagram', *Philosophical Transactions of the Royal Society of London, Series A* (332).
S. Kenna and S. Ross (1995). 'Networking in the Humanities.' London, etc: Bowker Saur.
E. Keogh, S. Lonardi and B.Y.-c. Chiu (2002). 'Finding Surprising Patterns in a Time Series Database in Linear Time and Space.' Presented at Eighth

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
D.V. Khmelev (2000). 'Disputed Authorship Resolution through Using Relative Entropy for Markov Chains of Letters in Human Language Texts', *Journal of Quantitative Linguistics* 7 (3): 201-207.
D.V. Khmelev and F.J. Tweedie (2001). 'Using Markov Chains for Identification of Writers', *Literary and Linguistic Computing* 16 (4): 299-307.
G. King (1997). *A Solution to The Ecological Inference Problem: Reconstructing Individual Behavior From Aggregate Data*. Princeton: Princeton University Press.
G. King and L. Zeng (2001). 'Explaining Rare Events in International Relations', *International Organization* 55 (3): 693-715.
S. King (1992). 'Record Linkage in a Protoindustrial Community', *History and Computing* 4 (1): 27-33.
S. King (1994). 'Multiple-source Record Linkage in a Rural Industrial Community, 1680-1820', *History and Computing* 6 (3): 133-142.
W. Kintsch (2003). 'On the notion of theme and topic in psychological peocess models of text comprehension', in: W.v. Peer, *Parsing for the theme. A computer based approach*. Amsterdam, Philadelphia: John Benjamins Publishing, 158-170.
M.G. Kirschenbaum (2002). 'Editor's Introduction: Image-based Humanities Computing', *Computer and the Humanities* 36: 3-6.
E. Klijn and Y.D. Lusenet (2002). ' In the Picture. Preservation and Digitisation of European Photographic Collections.' Amsterdam: European Commission on Preservation and Access.
E. Klijn and L. Sesink (2003). 'SEPIA Working Group on Descriptive Models and Tools.' Presented at SEPIA Conference 2003, Helsinki.
A.K. Knowles (2002). 'Introducing Historical GIS', in: A.K. Knowles, *Past Times, Past Place. GIS for history*. Redlands, CA.: ESRI Press.
M. Kobialka (2002). ''Can there be such a thing as a postmodern archive?'', in: J. Frow, *The New Information Order and the Future of the Archive*. Institute for Advanced Studies in the Humanities - The University of Edinburgh.
J. Kok (1997). 'Youth Labour Migration and its Family Setting, the Netherlands 1850-1940', *History of the Family. An International Quarterly* 2: 507-526.
J. Komlos and M. Artzrouni (1994). 'Ein Simulationsmodell der Industriellen Revolution', *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* 81 (3): 324-338.
I. Koprinska and S. Carrato (2001). 'Temporal Video Segmentation: a Survey', *Signal Processing: Image Communication* 16 (477-500).
J.M. Kousser (2001). 'Ecological Inference from Goodman to King', *Historical Methods* 34 (3): 101-126.

I. Kreft and J.d. Leeuw (1998). *Introducing Multilevel Modelling*. London: Sage.
G. Kristiansson (2000). 'Building a National Topographic Database'. Derived from the World Wide Web: http://www.geog.port.ac.uk/hist-bound/project_rep/NAD_more_info.htm.
H. Kropaç (1997). 'Electronical Documentation vs. Scholarly Editing?' Derived from the World Wide Web: http://www.hist.uib.no/acohist/kropac/kropac.htm.
G.P. Landow (1996?). 'Hypertext, Scholarly Annotation, and the Electronic Edition'.
E. Lecolinet, L. Robert and F. Role (2002). 'Text-image Coupling for Editing Literary Sources', *Computers and the Humanities* 36: 49-73.
Q. Li and R. Reuveny (2003). 'Economic Globalization and Democracy: an Empirical Analysis', *British Journal of Political Science* 33 (1): 29-54.
G. Lind (1994). 'Data Structures for Computer Prosopography', in: H.J. Marker and K. Pagh, *Yesterday. Proceedings from the 6th international conference Association of History and Computing, Odense 1991*. Odense: Odense University Press, 77-82.
C. Litzenberger (1995). 'Computer-based Analysis of Early-modern English Wills', *History and Computing* 7 (3): 143-151.
J. Ljungberg (2002). 'About the Role of Education in Swedish Economic Growth, 1867-1995', *Historical Social Research* 27 (4): 125-139.
M. Louwerse (2003). 'Computational retrieval of themes', in: W.v. Peer, *Parsing for the theme. A computer based approach*. Amsterdam, Philadelphia: John Benjamins Publishing, 189-212.
M. Louwerse and W.v. Peer (2002). 'Thematics. Interdisciplinary Studies.' in *Converging Evidence in Language and Communication Research*. Amsterdam, Philadelphia: John Benjamins.
K.A. Lynch and J.B. Greenhouse (1994). 'Risk Factors for Infant Mortality in Nineteenth-century Sweden', *Population Studies* 48 (1): 117-135.
C. Martindale and D. McKenzie (1995). 'On the Utility of Content Analysis in Author Attribution: *The Federalist*', *Computers and the Humanities* 29 (4): 259-270.
E. Mawdsley, N. Morgan, L. Richmond, et al. (1990). 'History and Computing III. Historians, Computers and Data. Applications in Research and Teaching.' Manchester, New York: Manchester University Press.
H.J. McCammon (1999). 'Using Event History Analysis in Historical Research: with Illustrations from a Study of the Passage of Women's Protective Legislation', in: L.L. Griffin and M. van der Linden, *New Methods for Social History (International Review of Social History, Supplement 6)*. Cambridge: CUP, 33-56.

W. McCarty (1999). 'Humanities Computing as Interdiscipline'. Derived from the World Wide Web: http://www.kcl.ac.uk/humanities/cch/wlm/essays/inter/.
W. McCarty (2001). 'Looking Through an Unknown, Remembered Gate: Millennial Speculations on Humanities Computing'. Derived from the World Wide Web: http://www.kcl.ac.uk/humanities/cch/wlm/essays/victoria/.
B.H. McCormick, T.A. Defanti and M.D. Brown (1987). 'Visualization in Scientific Computing - A Synopsis', *Computer Graphics & Application* 7: 61-70.
L.J. McCrank (2002). *Historical Information Science. An Emerging Discipline*. Medford, New Jersey: Information Today.
J. McGann (1991). 'What is Critical Editing?' *TEXT: Transactions of the Society for Textual Scholarship* 5: 15-29.
J. McGann (1992). *A Critique of Modern Textual Criticism*. Charlottesville: UP of Virginia.
J. McGann (1995). 'The Rationale of HyperText'. Derived from the World Wide Web: http://www.iath.virginia.edu/public/jjm2f/rationale.html.
J. McGann (2002). 'Dialogue and Interpretation at the Interface of Man and Machine. Reflections on Textuality and a Proposal for an Experiment in Machine Reading', *Computer and the Humanities* 36: 95-107.
T. McIntosh (2001). 'Urban Demographic Stagnation in early Modern Germany: a Simulation', *Journal of Interdisciplinary History* 31 (4): 581-612.
J.C. Meister (2003). 'Parsing for the theme. A computer based approach', in: W.v. Peer, *Thematics. Interdiciplinary Studies*. Amsterdam, Philadelphia: John Benjamins Publishing, 407-431.
T. Merriam (2002). 'Linguistic Computing in the Shadow of Postmodernism', *Literary and Linguistic Computing* 17 (2): 181-192.
R. Metz (1988a). 'Ansätze, Begriffe und Verfahren der Analyse Ökonomischer Zeitreihen', *Historical Social Research* 13 (3): 23-103.
R. Metz (1988b). 'Erkenntnisziele Zeitreihenanalytischer Forschung', *Historical Social Research* 13 (3): 6-22.
R. Metz (1993). 'Probleme der Statistischen Analyse langer historischer Zeitreihen', *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* 80 (4): 457-486.
H. Mielants and E. Mielants (1997). 'The Importance of Simulation as a Mode of Analysis: Theoretical and Practical Implications and Considerations', *Belgisch Tijdschrift voor Nieuwste Geschiedenis* 27 (3-4): 293-322.
C. Monroy, R. Kochumman, R. Furuta, et al. (2002a). 'Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Documents', in: K. Börner and C. Chen, *Visual Interfaces to Digital Libraries. Lecture Notes in Computer Science*. Heidelberg: Springer-Verlag, 39-49.

C. Monroy, R. Kochumman, R. Furuta, et al. (2002b). 'Visualization of Variants in Textual Collations to Analyze the Evolution of Literary Works in the Cervantes Project.' *ECDL 2002*: 638-653.

R.J. Morris (1995). 'Death, Property and the Computer - Strategies for the Analysis of English Wills in the First Half of the Nineteenth Century', in: P. Teibenbacher, *The Art of communication. Proceedings of the Eight International Conference of the Association for History and Computing, Graz, Austria, August 24-27, 1993*. Graz: Akademische Druck- iund Verlaganstalt, 164-178.

F. Mosteller and D.L. Wallace (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading: Addison-Wesley.

V. Mueller-Benedict (2000). 'Confirming Long Waves in Time Series of German Student Populations 1830-1990 Using Filter Techniques and Spectral Analysis', *Historical Social Research* 25 (3-4): 36-56.

C. Mullings (1996). 'New Technologies for the Humanities.' London: Bowker Saur.

J.W. Nibbering and J. DeGraaff (1998). 'Simulating the Past: Reconstructing Historical Land Use and Modeling Hydrological Trends in a Watershed Area in Java', *Environment and History* 4 (3): 251-278.

H. Obinger and U. Wagschal (2001). 'Families of Nations and Public Policy', *West European Politics* 24 (1).

B.S. Okun (1995). 'Distinguishing Stopping Behavior from Spacing Behavior with Indirect Methods', *Historical Methods* 28 (2): 85-96.

J. Oldervoll (1992). 'Wincens, a Census System for the Nineties?' in: *Eden or Babylon? On Future Software for Highly Structured Historical Sources*. St. Katharinen: Max-Planck-Insitut für Geschichte, Göttingen, 37-52.

J. Oldervoll (1994). 'Why don't We All use dBase?' in: H.J. Marker and K. Pagh, *Yesterday. Proceedings from the 6th international conference Association of History and Computing, Odense 1991*. Odense: Odense University Press, 135-139.

B. Opheim (2000). 'Political Networks and Factions: Online Prosopography of Medieval Scandinavian Sagas', *History and Computing* 12 (1): 43-57.

S.E. Ostrow (1998). 'Digitizing Historical Pictorial Collections for the Internet'. Derived from the World Wide Web: http://www.clir.org/pubs/ostrow/pub71.html.

T. Ott and F. Swiaczny (2001). *Time-integrative Geographic Information Systems: Management and Analysis of Spatio-temporal Data*. Heidelberg: Springer Verlag.

W. Ott (2002). 'Textual Criticism / Scholarly Editing'. Derived from the World Wide Web: http://www.rdg.ac.uk/languages/allcach/textual.html.

M. Overton (1995). 'A Computer Management System fro Probate Inventories', *History and Computing* 7 (3): 135-142.

A.C. Pacek and B. Radcliff (2003). 'Voter Participation and Party-Group Fortunes in European Parliament Elections, 1979-1999: a Cross-national Analysis', *Political Research Quarterly* 56 (1): 91-95.

M.E. Palmquist, K.M. Carley and T.A. Dale (1997). 'Applications of Computer-Aided Text Analysis: Analyzing Literary and Nonliterary Texts', in: C.W. Roberts, *Text Analysis for the Social Sciences*. New Jersey: Erlbaum, 171-189.

D. Parker (2001). 'The World of Dante: a Hypermedia Archive for the Study of the *Inferno*', *Literary and Linguistic Computing* 16 (3): 287-297.

C. Plaisant, B. Milash, A. Rose, et al. (1996). 'LifeLines: Visualizing Personal Histories.' Presented at CHI'96, Vancouver, BC.

A. Prescott (1997). 'The Electronic Beowulf and Digital Restoration', *Literary and Linguistic Computing* 12: 185-195.

C.T.D. Price, F.G. O'Brien, B.P. Shelton, et al. (1999). 'Effects of Salicylate and Related Compounds on Fusidic Acid MICs in Staphylococcus Aureus', *Journal of Antimicrobial Chemotherapy* 44: 57-64.

S.A. Raaijmakers (1999). 'Woordsoorttoekenning met Markov-modellen', in: *Jaarboek van de Stichting Instituut voor Nederlandse Lexicologie, overzicht van het jaar 1998*. 82-90.

L.E. Raffalovich (1999). 'Growth and Distribution: Evidence from a Variable-parameter Cross-national Time-series Analysis', *Social Forces* 78 (2): 415-432.

L.E. Raffalovich and D. Knoke (1983). 'Quantitative Methods for the Analysis of Historical Change', *Historical Methods* 16 (4): 149-154.

D. Reher and R. Schofield (1993). 'Old and New Methods in Historical Demography.' Oxford: Clarendon.

A. Reid (2001). 'Neonatal Mortality and Stillbirths in early Twentieth Century Derbyshire, England', *Population Studies* 55 (3): 213-232.

K.F.J. Reinders (2001). *Feature-based Visualization of Time-dependent Data*. Delft: Diss. TU Delft.

A. Renear, E. Mylonas and D. Durand (1993). 'Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies'. Derived from the World Wide Web: http://www.stg.brown.edu/resources/stg/monographs/ohco.html.

A. Ritschl (1998). 'Reparation Transfers, The Borchardt Hypothesis and the Great Depression in Germany, 1929-32: a Guided Tour for Hard-headed Keynesians', *European Review of Economic History* 2 (1): 49-72.

P.M.W. Robinson, W. Gabler and H. Walter (2000). 'Making Texts for the Next Century.'

G. Rockwell (1999). 'Is Humanities Computing an Academic Discipline?'

J. Rudman, D.I. Holmes, F.J. Tweedie, et al. (1997). 'The State of Authorship Attribution Studies: (1) The History and the Scope; (2) The Problems -- To-

wards Credibility and Validity.' Derived from the World Wide Web: http://www.cs.queensu.ca/achallc97/papers/s004.html.

R. Ruusalepp (2000). 'Multiple-Source Nominal Record Linkage: An Interactive Approach with Κλειω', in: P. Doorn, *Data Modelling Modelling History. Proceedings of the XI International Conference of the Association for History and Computing, Moscow, August 1996*. Moscow: Moscow University Press, 320-332.

J.A. Rydberg-Cox, R.F. Chavez, D.A. Smith, et al. (2002). 'Knowledge Management in the Perseus Digital Library', *Ariadne* 25.

J. Schellekens (1995). 'Illegitimate Fertility Decline in England, 1851-1911', *Journal of Family History* 20 (4): 365-377.

C. Schonhardt-Bailey (1998). 'Parties and Interests in the 'Marriage of Iron and Rye'', *British Journal of Political Science* 28 (2): 291-332.

R. Schor (1996). *Histoire de l'immigration en France de la fin du XIXe siècle à nos jours*. Paris: Armand Colin.

A.T. Schreiber, B. Dubbeldam, J. Wielemaker, et al. (2001). 'Ontology-based Photo Annotation', *IEEE Intelligent Systems* 16 (66-74).

S. Schreibman (2002). 'Computer-mediated Texts and Textuality: Theory and Practice', *Computers and the Humanities* 36 (3): 283-293.

A. Schuurman and G. Pastoor (1995). 'From Probate Inventories to a Data Set for the History of the Consumer Society', *History and Computing* 7 (3): 126-134.

S. Scott, S.R. Duncan and C.J. Duncan (1998). 'The Origins, Interactions and Causes of the Cycles in Grain Prices in England, 1450-1812', *Agricultural History Review* 46 (1): 1-14.

D.A. Smith (2002). 'Detecting Events with Date and Place Information in Unstructured Text'. Derived from the World Wide Web: www.perseus.tufts.edu/Articles/datestat.pdf.

D.A. Smith, J. Rydberg-Cox and G.R. Crane (2000). 'The Perseus Project: a Digital Library for the Humanities', *Literary and Linguistic Computing* 15 (1): 15-25.

B.K. Song (2002). 'Parish typology and the Operation of the Poor Laws in early Nineteenth-Century Oxfordshire', *Agricultural History Review* 50 (2): 203-224.

S.J. South (1999). 'Historical Changes and Life Course Variation in the Determinants of Premarital Childbearing', *Journal of Marriage and the Family* 61 (3): 752-763.

W.A. Speck (1994). 'History and Computing: Some Reflections on the Past Decade', *History and Computing* 6 (1): 28-32.

R. Spree (1997). 'Klassen- und Schichtbildung im Medium der Privaten Konsums: Vom Späten Kaiserreich in die Weimarer Republik', *Historical Social Research* 22 (2): 29-80.

D.J. Staley (1998). 'Designing and Displaying Historical Information in the Electronic Age', *Journal of the Association for History and Computing* 1 (1).
D.J. Staley (2003). *Computers, Visualization, and History: How New Technology Will Transform Our Understanding of the Past*. Armonk, N.Y.: M.E. Sharpe.
L.L. Stewart (2003). 'Charles Brockden Brown: Quantitative Analysis and Literary Interpretation', *Literary and Linguistic Computing* 18 (2): 129-138.
W. Stier (1989). 'Basic Concepts and new Methods of Time Series Analysis in Historical Social Research', *Historical Social Research* 14 (1): 3-24.
M. Thaller (1980). 'Automation on Parnassus. CLIO - A databank oriented system for historians', *Historical Social Research* 15: 40-65.
M. Thaller (1987). 'Methods and Techniques of Historical Computation', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 147-156.
M. Thaller (1989). 'The Need of a Theory of Historical Computing', in: P. Denley, S. Fogelvik, and C. Harvey, *History and Computing II*. Manchester, New York: Manchester University Press, 2-11.
M. Thaller (1993a). 'Historical Information Science: Is There such a Thing? New Comments on an old Idea', in: T. Orlandi, *Seminario Discipline Umanistiche e Informatica. Il Problema dell' Integrazione*. Roma.
M. Thaller (1993b). 'What is 'Source Oriented Data Processing?'; What is a 'Historical Information Science?'', in: L.I. Borodkin and W. Levermann, *Istoriia i comp'iuter. Novye informationnye tekhnologii v istotricheskikh issledovanii akh i obrazovanii*. St. Katharinen, 5-18.
M. Thaller (1996). 'Digital Manuscripts: Editions v. Archives'. Derived from the World Wide Web: http://www.hit.uib.no/allc/thaller.pdf.
S. Thernstrom (1973). *The other Bostonians : Poverty and Progress in the American Metropolis, 1880-1970*. Cambridge, Mass.: Harvard University Press.
P. Tilley and C. French (1997). 'Record Linkage of Nineteenth-century Census Returns. Automatic or Computer Aided?' *History and Computing* 9 (1-3): 122-133.
E. Tufte (1983). *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.
F.J. Tweedie, S. Singh and D.I. Holmes (1996). 'Neural Network Applications in Stylometry: The Federalist Papers', *Computers and the Humanities* 30: 1-10.
E. Urbina, R.K. Furuta, A. Goenka, et al. (2002). 'Critical Editing in the Digital Age: Information and Humanities Research', in: J. Frow, *The New Information Order and the Future of the Archive*. Institute for Advanced Studies in the Humanities - The University of Edinburgh.

E. Vanhoutte (1999). 'Where is the editor? Resistance in the creation of an electronic critical edition', *Human IT. Tidskrift för studier av IT ur ett humanvetenskapligt perspektiv* 1.
S.v.d. Velden and P.K. Doorn (2001). 'The Striking Netherlands: Time Series Analysis and Models of socio-economic Development and Labour Disputes, 1850-1995', *Historical Social Research* 26: 222-243.
J.E. Vetter, J.R. Gonzalez and M.P. Gutmann (1992). 'Computer-Assisted Record Linkage Using a Relational Database System', *History and Computing* 4 (1): 34-51.
J. Viscomi (2002). 'Digital Facsimiles: Reading the William Blake Archive', *Computers and the Humanities* 36: 27-48.
C.C. Webb and V.W. Hemingway (1995). 'Improving Access: A Proposal to Create a Database for Probate Records at Borthwick Institute', *History and Computing* 7 (3): 152-155.
G. Welling (1993). 'A Strategy for Intelligent Input Programs for Structured Data', *History and Computing* 5 (1): 35-41.
G. Welling (1998). *The Prize of Neutrality. Trade relations between Amsterdam and North America 1771-1817. A study in computational history*. Hilversum: Verloren.
G.M. Welling (1992). 'Intelligent Large-scale Historical Direct-data-entry Programming', in: J. Smets, *Histoire et Informatique. Actes du Congrès. Ve Congrès 'History & Computing' 4-7 Septembre 1990 à Montpellier*. Montpellier, 563-571.
J.J.v. Wijk and E.v. Selow (1999). 'Cluster and Calendar-based Visualization of Time Series Data.' Presented at Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99), October 25-26, 1999.
I. Winchester (1970). 'The Linkage of Historical Records by Man and Computer', *Journal of Interdisciplinary History* 1: 107-124.
R.L. Woods (1987). 'Skills for Historians: Getting Something Done with a Computer', in: P. Denley and D. Hopkin, *History and Computing*. Manchester: Manchester University Press, 205-210.
M. Woollard (1999). 'Introduction: What is History and Computing? An Introduction to a Problem', *History and Computing* 11 (1-2): 1-8.
E.A. Wrigley (1973). *Identifying People in the Past*. London.
K. Yamaguchi (1991). *Event History Analysis*. Beverly Hills / New York: Sage.
D. Zeldenrust (2003). 'Picture the Past, the Use of Documentary Photographic Images in Historical Research.' Presented at SEPIA Conference 2003, Helsinki.
Z. Zhao (1996). 'The Demographic Transition in Victorian England and Changes in English Kinship Networks', *Continuity and Change* 11 (2): 243-272.

Acknowledgements

This study has been made possible thanks to the generous co-operation of the Vakgroep Informatica en Informatiekunde of the University of Utrecht, and the Faculteit Letteren of the Radboud University Nijmegen, who allowed Leen Breure and Onno Boonstra to work for the NIWI project on past, present and future of historical information science.

Parts of this study have been discussed at various occasions with a wide range of specialists in the field. We would like to thank all those that were so kind to provide us with new information or with critical remarks during these sessions. In particular, we would like to express our gratitude to:

- the participants of the workshop 'Past, present & future of historical information science', 9 August 2003, International AHC Conference, Tromsø, Norway;
- the participants of the workshop 'Historical information science', 24 October 2003, Amsterdam;
- the participants at the VGI symposium 'Past, present & future of historical information science', 17 November 2003, Amsterdam;
- dr. Peter Boot, Constantijn Huygens Instituut, Den Haag
- prof. dr. Hans Bennis and drs. Edwin Brinkhuis, Meertens Instuut, Amsterdam;
- dr. Donald Haks and dr. Rik Hoekstra, Instituut voor Nederlandse Geschiedenis, Den Haag;
- dr. Henk Wals, International Institute for Social History, Amsterdam;
- dr. Hans Voorbij and prof. dr. Mark Overmars, Instituut Informatica en Informatiekunde, Universiteit Utrecht;
- dr. Karina van Dalen, Afdeling Neerlandistiek, Nederlands Instituur voor Wetenschappelijke Informatiediensten, Amsterdam;
- the National Archives, Den Haag;
- dr. Truus Kruyt, Instituut voor Nederlandse Lexicologie, Den Haag;
- dr. Martin Bossenbroek, drs. Marco de Niet, Mr. Margariet Moelands, Koninklijke Bibliotheek, Den Haag;
- prof. dr. Eep Talstra, Faculteit Theologie, Vrije Universiteit Amsterdam;
- dr. George Welling, Vakgroep Alfa-informatica, Rijksuniversiteit Groningen

The first draft of this report was discussed on an international meeting, held in 'De Sparrenhorst', Nunspeet, 13-15 February 2004, by the following experts:

- prof. dr. Manfred Thaller, Historisch-Kulturwissenschaftliche Informationsverarbeitung, Universität zu Köln, Germany;
- prof. Gunnar Thorvaldsen, Norwegian Historical Data Centre, Faculty of Social Sciences, University of Tromsø, Norway;
- prof. Leonid Borodkin, Historical Informatics Lab, Moscow State University, Russia;
- dr. Matthew Woollard, Arts and Humanities Data Service, History, University of Essex, UK;
- dr. Jan Oldervoll, Historisk Institutt, University of Bergen, Norway;
- prof. dr. Henk Koppelaar, Department of Information Technology and Systems, Technical University Delft;
- prof. dr. Martin Kersten, Instituut voor Wiskunde en Informatica, Amsterdam University;
- prof. dr. Jaap van den Herik, Department of Computer Science, Universiteit Maastricht;
- prof. Bob Morris, Economic and Social History, University of Edinburgh, UK;
- prof. Dr. Ingo H. Kropac, Institut für Informationsverarbeitung in den Geisteswissenschaften, Karl-Franzens-Universität Graz, Austria;
- Alan Morrison, Oxford Text Archive / AHDS literature, language and linguistics, Oxford University, UK;

We are especially grateful to Matthew Woollard, who not only made many wise and useful comments on the first draft of this text, but who also checked and corrected our broken English.