Data bases v. critical editions

Thaller, Manfred

COMPUTER SECTION

Data Bases v. Critical Editions

*Manfred Thaller**

1. Editions and Computers

Historical studies as we know them today, are the result of a long development. Many of the phases of this history of History could have taken place differently, without changing the outcome into something, we would not recognize any more. Indeed, like happiness, history still means very different things to different people — what could scarcely be observed better anywhere else than at such a large-scale meeting of historians as the congress at Kalamazoo. Subjects and methods of historians are so different, that, indeed, some of us may find themselves much more closely related to our colleagues from other departments than to historians from other subdisciplines.

What distinguishes history as an academic discipline from purely speculative writing about the past, is its claim to present the times gone by in a way, which can be criticized not because of a particular literary taste, but because whatsoever has been said about an earlier epoch can be checked with the help of the sources quoted as the firm ground upon which a particular thesis has been built. So almost the only development of the last few hundred years, which could not have been left out, without changing historical studies into something which we would not recognize any more, is the process that let from the criticism applied by L. Valla to the donation of Constantine to the critical editions of today. Indeed, it would be hard to imagine, what medieval history would be like in the United States, if it would not be possible to have significant documents of the transatlantic past on ones' shelves, but one would have to cross the Atlantic every time, when one intends to consult a charter, the only copy still residing in the coffers of some noble household.

* Address all communications to: Manfred Thaller, Max-Planckinstitut für Geschichte, Hermann-Föge-Weg 11, D-3300 Göttingen.

Historians have always been lovers of the printed word; while first loves tend to linger on, however, the scarcely stay exclusive: and particularly during the more recent decades, historians became increasingly interested in other types of sources, too. Think of collections of medieval pictures, of administrative relicts like account books and the like. Quite a few of these new types of source have put up considerable obstacles for traditional editing techniques: few of us would doubt, e.g., that the editions of *necrologia* printed in the 19th century, are of much less value today, than editions of the same time dealing with material which contains more substantive text, like charters. What all the types of source material which are and have been used, have in common, is scarcely more than that the constitute informations and a substantial proportion of the life of every historian is spent in trying to retrieve the information needed from a vast collection of potentially relevant corpora.

We, today, live in what has been called the »age of informations are going through what is supposed to be the information revolutions Small wonder, that from the very first conferences on the applicability of computer technology to the Humanities onwards, the question of how to use a computer to retrieve quickly a needed item of information out of a vast heap thereof has been a central concern of historians (1). And let me just remind briefly, that the very first of them all, the *Index Thomisticus* of Padre Busa (2) dealt with a task which is almost an editorial one: how to make the writings of Aquinas most readily available to the academic community.

So we would expect, that the computer would have a significant impact upon a discipline which is so intimately related to the retrieval and the handling of information. Has it? One can scarcely doubt, particularly if looking at the other »computer sessions« of this congress, that data processing *has* made its impact. Historians do not only *read* books; they also write them. And whatsoever makes the life of another writer easier improves also the lot of the author of a historical study. Still one might wonder, if this impact is not a little bit superficial. Writing a book, we are presenting a result: the significance of the scholarly edition is with the conducting of research, not the publication of its findings. This is not to doubt the impact of computers: historical studies, which have produced their results with the help of statistical or other software, exist in ever increasing numbers.

There does exist a strange tendency, however. Most historians, which justify the use of computer technology in the introduction of a book or project report, do so by hinting at the very large amount of source material, which exists for their topic of research. This, it is said, could not be controlled without using either statistical abstraction, or the ability of the computer to perform searches quickly in vast heaps of material, or both.

When one compares the actual sources, upon which the findings are based, one sometimes gets the impression, however, that the traditional historian has a tendency to use more sources, than his computer using colleague. These sources are analyzed and used much less intensively of course: appearing in single quotations, or used just as additional illustration for a subtle point. The more heavily a study uses the computer, however, the more it concentrates usuallly upon a single central source, or at least upon a very precisely defined corpus of similar sources, which than are analyzed much more intensively, as, e.g., by a statistical type of presentation. The reason for this is, that in some ways the computer using historian is in the situation of the days before the appearance of the scholarly edition: before you can use the power of computational tools on a source, you have to transcribe it completely -- as you had to, when all a historian could base his findings upon was, whatsoever he could copy by pen from the material made accesible to him by the benevolence of a local archivist.

This situation exists, though in the last twenty years there have probably been very few funding proposals which did not mention it as an attractive side effect of the respective project, that it would make particularly important source material more easily accesible to the academic community, if that materia would be turned into machine readable form. There exist at least three reasons, these promises seem never to have been kept:

1) Firstly, machine-readable data abstract traditionally very much from the source; very much more so, in most cases, than a printed edition. Very often it seems to be more easy to go through the original source than to try and understand, what precisely a given convention of turning a handwritten list into a statistical code actually implies; just besides the point, that, even if the machine-readable data are well documented, this process of abstraction goes so far, that historians will quite instinctively doubt if they are not led astray by a misunderstanding of the material commited by the person who entered it into the computer at first hand.
2) Secondly, while software solutions, which avoid these problems, allowing the entry of a source more or less verbatim, exist, these software solutions tend to be purely local; so, if you do not have the specific program, it is almost impossible to use the machine-readable material. And even if you have it, the conventions which are used for the structuring of the data are inherently much more complex than the conventions with traditional, numerically coded material. So, even if you have the software available, it can be so cumbersome to understand how, precisely, a given portion of the text has been turned into a machine readable item, that once again it may be easier to turn to the original source than trying to understand how it has been prepared.
3) The third question is finally a question of accessibility. Let's assume we

have a data bank, which contains short information on every person in a given region of, say, the 9th century. At first glance this seems to be a tool, which should be bound to fascinate everybody working about this time: you find a name somewhere in a document and within seconds you know, where this name occurs again. Unfortunately, before these seconds start, there are sometimes a few thousand miles to cross, to reach this data base system; and, even if you are allowed and able to access it via a network, this still implies to learn a specific command language for these single enquiry.

Today, there exist at least two approaches to improve upon this situation:

1) On the one hand, thanks to the efforts of Theodore Brunner and the *Packard Humanities Institute,* there exists the instrument of the *Thesaurus Linguae Graecae,* holding every Greek text that has been written before the year 600; available on the *Ibycus Scholarly Computer* (3), a dedicated work station type of machine, which could very well be afforded by a research institution.
2) On the other hand, the *Research Library Group* is currently a model for improved access to certain types of historical information on the public networks, with the project of the *Medieval and Early Modern Data Bank* (4), which will be presented at this congress. Though we admire these approaches very much and would not dare to dispute their enormous merits, we would like to point out, that both of them geared towards the solution of relatively specific problems, while the general situation we described initially can be improved upon only, if we provide additional tools for research.

In the first case we think, that the situation of the *Thesaurus Linguae Graecae* is unique, because it is dedicated to the treatment of a large, but limited corpus: if you count the bits and pieces, all the surviving texts of ancient Greek will probably add up to less, than the holdings of even a single small local archive for the 14th century, say in Flandres. We think therefore, that while it is scarcely possible to overestimate the usefulness of a tool like the TLG, it is a model of research, which can not be applied to other areas, where the sheer volume of the relevant source material is at another magnitude. Even the most obscure text of Greek, which has survived from the, say, 3rd century, is relevant for a relatively high percentage of the section of the academic community dealing with ancient Greek at all. The number of researchers who would be interested, say, in the account book of a particular monastery is incomparably smaller and it would be out of the question to organize a dedicated research setup to create a workstation-type database to further research upon a specific town, the content of which would be interesting only to a small number of local historians. Still such account books have been published and many local

archives have their own series of source editions; which are used, not frequently, but over a long time by researchers who casually browse through the collection of sources their institution holds to find examples of parallels to or deviations from the specific pattern of events they are doing research upon.

In the second case, our argument has to be more subtle; or not subtle at all, because after all it seems to us to be a question of money. While we agree, of course, that there are central sources, which should be made available centrally — such as the ones which will be demonstrated at this congress -- we have the strong impression, that the vast majority of historical sources is of a type, which will be so infrequently consulted and takes up such a large amount of storage, that it will be not so economical to keep those materials on-line.

2. The Concept of a Historical Workstation

To solve these problems and continuing after previous projects, at the *MaxPlancklnsTirut für Geschichte* a concept has been developed for a multi-purpose workstation, which should be dedicated to the kind of problems which are specific for academic work in historical research. In the following paragraphs we would like to describe this concept and report on the current state of its realization. After this, we will focus on one of it's aspects, the question, how machine readable editions could be provided; how, that is, within such a larger concept the difficulties described above could be solved. To prove, that this is not just a concept, but a working project, we will then shortly introduce five important contributions towards such a collection of machine readable editions, which are available already for public distribution: each of them will in turn be described in greater detail in the remaining papers of these two sessions, so we focus upon the way in which these public domain data bases fit into the general design.

The technical definition of a historical workstation, as we employ the term, is:

A desktop computer, which

- has access to data base management software, which is able to administrate very different structures of information, allowing to put into a data base arbitrarily large collections of sources, keeping as much information of the original and applying as little coding, as is economically feasible for the project producing the data base,
- has access to a set of data bases, which contain background knowledge specific to historical research,
- has access to a large number of read-only data bases, being equivalent

to traditional printed editions of source material,

- contains sufficient Artificial Intelligence subsystems, to make the interaction between the forementioned capabilities transparent to the user,
- has a very highly integrated interface between the data base management systems mentioned and a desk-top publishing system and
- a similar interface to statistical software.

What this technical definition describes shall allow the following working situation for a historian. Let's assume, we are working on a history of a particular town in the fourteenth century. First of all, we simply consult all the machine readable sources available to us for references to that town; much as we would do browsing through the large printed collections of sources. When we detect portions of source material, which are related to our subject, we download it into a kind of »private data base«, which, unlike the machine readable sources distributed as part of the workstation, can be updated freely. (You change your notes quite frequently; you do not usually scribble unto the pages of editions, however.)

Parallely to that use of material available in this »edited« form, you start to prepare further entries into your data base, made up of unpublished source material -- if you consider your source important enough, you prepare it at the same time in a form, which might become available later as another machine readable edition, accredited intellectually to you.

During that process, you will encounter numerous items of information, which seem to be less than completely clear to you: names of persons, which you think might be identical to names of persons you are vaguely familiar with; names of currencies where you are not sure what exchange value they had at the time; items with a chronological bearing you have to rectify. In such cases you are able on the one hand to look these bits of information up in the specialized knowledge bases of your workstation. What is more important, however, is, that you can enter these items into your data base just as words and tell the system to remember, that it shall link them to the expert knowledge residing in the background: so, if this background knowledge becomes updated, because somebody discovered, that there exists more information about a specific person, an exchange rate was so far simply understood wrongly or a particular saint was related to another day in the bishopric about which you are working, these new discoveries are immediately integrated into your private data base.

3. The Reality of a Historical Workstation

How realistic is it, that we will see such a system to become workable in the foreseeable future? We seriously doubt, that there exists any country, where funding for the Humanities would be so ample, that a system like the one we described, could be realized as one, monolithically funded project. So the outline given above is not a »Workstation project«: it is a medium term guideline according to which we try to organize our ongoing research, which, generally is devoted to the improvement of the methodological tools available for historical research. What the *MaxPlanck-Institut für Geschichte* attempts, is to create a loose network of differently funded projects, which agree about the basic sensibility of that model of future research and try to realize individual components of such an overall model, which are designed to fit ultimately together within such a design. As of now, the following projects have been realized or are in the stages described (5):

1) The original project, from which the further developments are derived started in 1978 at the *MaxPlanckInstitut für Geschichte*. It attempted to create a prototype of data base software that should be able to handle historical sources as close to the original as possible, and to provide services specific to the needs of historians in the following fields: information retrieval, report generation, preparation of loosely structured sources for statistical analysis, nominal record linkage, complex file mergers and a full-text retrieval system. This project, which became known as CLIO, has been realized on a UNIVAC mainframe. Building upon the experiences gained with that system, a complete reimplementation under the name of ΨΛΙΟ was started in 1986. A first version has been released in 1987, a second one immediately before this congress; the third which is due in August of this year will reimplement the full functional power of the prototype. The complexity of the command language needed of the system is roughly comparable to such command languages like SPSSX or SAS. While currently negotiations for a provision of the system with a completely menu oriented frontend are under way, it has always been a determined policy of the *Max-Planck-Institut* to provide software solutions for problems which could not be dealt with by commercial software and make these available to the community of historical researchers free or at nominal cost, rather than putting limited funds into the developments of frontends which are very beautiful but functionally insignificant. As a consequence the software runs currently on seven hardware / operating system combinations and can be used, specifically, on PCs as well as on mainframes.
2) This re-implementation has been supported by funding for distinct

modules, provided by the *Instituí für mittelalterliche Realienkunde,* Krems, Austria, the *LudwigBoltzmannlnstitut jür Historische Sozialwissenschajten*, Salzburg and the *Instituí for Anthropology* of Göttingen university.

3) In the spring of 1987 an agreement was reached with the *Istituto Lingüistica Computazionale* in Pisa, Italy. The two institutions agreed to port the Latin lemmatization system developed in Pisa into C, according to the portability standards used for the development of ΨΛΙΟ and to integrate it fully into the data base system, to enhance the usefulness of the full text retrieval components.
4) In the fall of 1987 IBM Germany provided the cooperating projects with a study contract to integrate optical disks into a Humanities workstation environment.
5) Since January 1988 a two year project, funded by the Austrian ministry of research is under way at the *Forschungsinstitut "fur Historische Grundwissenschaften* of the University of Graz, which shall be responsible for improved possibilities for record linkage and the treatment of prosopographical information.
6) In the spring of 1988 the *Volkswagen* foundation of Germany decided to support a three year project, with two aims in mind: on the one hand the development of Artificial Intelligence components, to improve the accesibility of expert knowledge within the overall design of a more general workstation shall be undertaken; on the other hand software dedicated to the improved exchangability of data between existing historical data bases and the faciliated loading of texts made machine readable by devices like the Kurzweil Data Entry machine has to be realized. This project is due to begin operation on June 1st of 1988.
7) Negotiations with a number of research institutions in different European countries, dealing with other technical sub-projects, are at various stages of realization.

What do this various contributions imply for practical work?
The basic decision to implement a new system of data base software reflects the opinion, that the information contained in historical sources has a number of properties, which are not so common — and therefore unsupported — in commercial environments. Some of them are fairly obvious: historical data bases do not have fields of fixed length; fields are frequently missing; fields contain more than one value. Some of them are very specific for historical research: to cope with the variation in historical spelling or with the intricacies of fluctuating non-decimal currency systems, has simply no relationship to the problems commercial software is dedicated to. A third class of problems, finally, borders upon questions,

which are currently being researched in information science, but not implemented in readily available software. Think of the classical problem of prosopography how to decide if a *Henricus erfurtensis* is a certain Henry originating from *Erfurt* or a Henry who already was known under what a little bit later became the surname of his family? Or of another one, which is not so frequently mentioned, but becomes very serious, when we discuss easily large scale data bases: if you ask a data base to provide you with all persons having been borne in a given political entity, will it be clever enough to look up on its own, at what time the geographical definition of that entity changed? More likely not. The first two classes of problems are handled by the original developments of the *MaxPlanckInstitut*; the third is part of the project just having been funded by the *Volkswagen* foundation.

The importance of the lemmatization components provided by the *Istituto Linguistica Computazionale* almost explains itself, we assume: it simply means, that looking for material on a given subject in a collection of texts, you are not forced anymore to memorize what kind of grammatical changes may happen to the words, when you are looking them up.

Providing optical disks is in some ways less spectacular: being able to hold many hundred megabytes on a cheap medium of exchange, which can be used with a standard PC, however, makes precisely the difference between a PC which administrates a single private base and one which allows to access a very large of machine readable sources.

Prosopographical information is a very substantial proportion of the holdings of any historical data base: indeed, the problem of how to overcome the differences between names spelled at wide variance, is probably the most glamorous one of all the difficulties faced by the computer using historian. So we probably do not have to point out further, why this contribution is a particularly important one.

Exchanging data between existing systems of historical data bases may seem almost irrelevant in comparison and a subject of purely technical relevance. It is not. Being able to exchange machine readable sources freely, irrespective of the conventions that have been used, when they were made machine readable in the first place, will in many cases be the difference between yet another »data graveyard« and a publicly available computer presentation of a text.

4. A Prototype for Presentation

To test the interaction between the various components of software and a number of machine readable sources which could be considered prototypes of the various kinds of »data bases as editions« envisaged, a optical disk has been prepared as a prototype for the congress at Kalamazoo, which is available at nominal cost (basically the price of the disk) for demonstration purposes. It contains:

1) The current version of the ΨΛΙΟ software and, integrated into it, a first version of the system for automatic lemmatization of the Latin language presented by Andrea Bozzi and Giuseppe Capelli (»A Latin Morphological Analyzer«).
2) A data base of roughly 100.000 persons, being a section of the data base presented by Maria Hasdenteufel-Roding and Dieter Geuenich. (»A Data Base for Research on Names and Groups of Persons in the Middle Ages«).
3) The data base on medieval pictorial sources presented by Gerhart Jaritz (»Finding the Signs; Pictures of Medieval Life«).
4) The collection of sources dealing with medieval migrations presented by Albert Müller (»Migration and Prosopography«).
5) The combined data base on prosopography together with the machine readable chartulary of Styria, as presented by Ingo Kropač (»Homo ex Machina: Prosopography and Chartularies«).

This being a prototype, which in regular intervals shall become reissued being enhanced by additional data bases is a *prototype* indeed: none of us assumes, that it is the final solution for the questions raised in the first parts of this paper. It can be a model, however, how data bases could become available with the same casual ease, with which printed editions can be consulted. As such it invites criticism; its successive editions will gain by being used; and it is an invitation to join the functioning cooperative network outlined above.

Notes

This paper tries to outline the common background of the various contributions to the two sessions on *data base oriented source editions* held at the 23rd International Congress on Medieval Studies in Kalamazoo, 5-8 May, 1988. Selected papers of the two sessions are published in a booklet *Data Base oriented Source Editions;* this booklet can be ordered from the author (M. Thaller).

(1) E.g. Stephan Thernstrom: *The Historian and the Computer*, In: Edmund A. Bowles, Ed.: *Computers in Humanistic Research,* Englewood Cliffs, NJ, 1967, 73-81, here: 78.

2) One can probably pay the respect due to this pioneer in no better way, than by pointing out, that already in the very first volume on Computing and the Humanities known to this author, his work has been described as historical: Roy Wisbey: *Computers and Lexicography,* In: Dell Hymes, Ed.: *The Use of Computers in Anthropology*, London, etc., 1965, pp. 216-234, here: 222-225.

3) On the *TLG* and *Ibycus* see recently: *Bits & Bytes Review* 1/7, June 1987, pp. 1-6.

4) The Research Library Group News 12, January 1987, pp. 8-10.

5) The following list is only partially complete, as our concept of a historical workstation« is by no means restricted to medieval history; as this presentation is made at a medievalists congress, however, we focus upon such developments, which are related to medieval studies.