# The historical workstation project: part 1

Thaller, Manfred

Historical Social Research, Vol. 16 — 1991 — No. 4, 51-61

# The Historical Workstation Project (1)

***Manfred Thaller****

Abstract: Workstations have been a particularly hot topic in recent discussions of computer technology. The author argues, that, while the additional computing power provided by them would be welcome within the historical disciplines, a truly »historical« workstation is not defined so much by the computing capacity provided, but by the tools and environments geared specifically towards historical research which could and should be available on such platforms. In describing how such modules could interlink - and how their creation can be organized - a common ground for the papers following in this issue is prepared.

## 1. Preliminary Considerations

The term »workstation« is usually employed for a specific class of computer, defined by the hardware capabilities it has. The precise definition of these capabilities is fairly vague: about the only agreement one notices, is, that a workstation is »more« than a PC. This agreement ***is*** vague indeed: while few people would consider a PC using the 80386 processor from Intel as a »workstation«, when it is running under MS-DOS, there is a tacit agreement among many, that this machine is the very prototype of a low-end workstation, as soon as it is controlled by one of the more powerful operating systems, or, putting it only slightly less general: by UNIX.

This close association of UNIX with the concept of a workstation, derives from the way workstations have been introduced: originally they come from an engineering or »hard« science environment, where considerable computing power and/or sophisticated hardware, like high resolution screens, were required to support individual researchers. The typical example of such an application would be the CAD (Computer Aided Design) techniques, asking for immediately visible responses to changes in complex visualisations. What does this have to do with UNIX? - Well,

* Address all communications to Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, 3400 Göttingen.

even the best hardware can be crippled beyond usefulness by a sufficiently badly designed operating system. So many problems arise, like the one that the kind of applications we talk about need lots of memory: which comes natural to UNIX and is always handled cripplingly by PC operating systems like MS-DOS, no matter how many embellishments one puts around them.

The point we wanted to make by this recapitulation of facts, which most of the readers of this paper will be familiar with already, is, that »workstations« as they occur today, are the result of a clearly defined working situation which shall be supported. One might even point out, that, while multilayered windows and cleverly handled mice may make a program considerably easier to use, there ubiquitiousness among the existing workstations is *not* derived from this user friendliness in general, but from the fact that such an user interface is >natural< for a computer, which is dedicated to graphical applications to begin with.

*The workstation* of today is therefore derived from assumptions about what is needed, to support the work situation of a specific class of users: engineers and »hard« scientists. It follows logically, that a *historical* or *Humanities* workstation should be derived from an analysis of the needs of a scholar from these disciplines.

Out of this conclusion a few years ago the project we describe here has been derived. Or, to repeat what this author has already said a number of times: the historical workstation »project«. The quotation marks around the word »project« should not be misread as a typical German preoccupation with obscure points of terminology, but as a hint to an important organizational aspect. Development work in computing is expensive; if we accept the idea, that a »historical workstation« is not just a workstation type of computer used by a historian, but a computer equipped with software specifically developed on the basis of an understanding, what specifically historical requirements are, we reach very soon the multi-man-year type of development project. Or: the type of project which is simply not feasible, funding possibilities for Humanities research being what they are. So, when we speak about a »project« in this paper, and in the ones it serves as introduction for, we do not mean a clearly defined group of people working at one institution for a specific development goal: we rather speak about a loosely organized group of research institutions and researchers, which are coordinating their various efforts to produce research designs, pieces of software and data bases, according to specifications allowing them to fit together as building blocks for a coherent working environment.

Historical Social Research, Vol. 16 — 1991 — No. 4, 51-61

## 2. Our Concept of a Workstation

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

*A 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 managment system mentioned and a desk-top publishing system and
- a similar interface to statistical software.
- All interpretation of information is so far based on transcribed (i.e. byte coded) texts. Such texts can, however, be linked to bit mapped images or portions thereof, making an image a candidate for processing as result of an arbitrarily complex query.

To give a more practical impression of what this concept stands for, we would like to describe the typical process of its being used:

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 into 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.

While some of the various types of data bases to be used with such a system will be described in the following contributions and/or the demonstrations at the Siegen conference, we would like to point specifically to two strategic decisions:

1) 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 available large scale data bases: if you ask a data base to provide you with all persons having been born 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.

2) For the initiated the strangest thing about the technical definition has probably been an omission: we just mentioned *a computer,* neither dwelling upon the great power of the Mac, nor explaining why we considered X Windows the way to go, nor even mentioning UNIX again in this more technical part of our presentation, though we did so

in our introductory remarks. This reflects the opinion of the project group, that a *historical* workstation has truly to be defined by its processing capabilities; not by the hardware it runs upon. So, while all software development is done in the programming language >C< and indeed under UNIX, portability for all solutions found, has always been one of the most important goals. As a result, most of the software discussed here, runs on about seven different operating systems, right from PC's under MS-DOS through to the super-minis of CONVEX and IBM's 3090 line of mainframes under VM/CMS (2) This has an important implication: we started from a strong bias for the work horse, rather than for the beautiful dancer. Data bases supported by the software described here go well beyond 200 MB, with three multi-GB ones currently in their concepualization phase. But: as prices currently go, most of our applications projects had to choose between really large disk capacities and good graphics capabilities; and with very few exceptions took the first decision. Obviously one has to make compromises with portability in some cases: features for cartography are available only with some of the versions of the system; the image processing components are currently available only under X Windows on machines which run IBM's AIX.

## 3. Building Blocks and Highlights

The concept of a Historical Workstation grew out of the experiences gained with a project named CLIO. Quite naturally the successor to that project, a DBMS known as Kteuo, still provides the backbone of the concerted attempts at software development. Many of the projects touched upon in the following paragraphs use software support libraries developed when a generalization of the original software started as an individual project, funded by the German *Volkswagen Foundation* (3)

Around this backbone developments started into a number of directions. Going from software currently being shipped, to software in advanced stages of development and further on to such, which is in earlier stages, some of the most important items are the following.

### 3.1 Kteuo- A DBMS

Currently version 3 of the basic DBMS software is being shipped, to about 200 research institutions and individual researchers.

Before we start to summarize it's highlights, an important reservation should be made: the development goals of the early years (since about 1986) have been to develop as quickly as possible as powerful a system as

feasible, with intentional neglect of the user interface. As a result, the version still being demonstrated at the Siegen conference is still being controlled by a command language, which is powerful, but definitely much more difficult to use than menu oriented or graphical interfaces as provided by modern commercial software. This will change as of version 4, due at the end of this year.

In a nutshell the system:

- is built around a newly developed data model, based upon the notion of semantic networks. It allows the handling of arbitrarily complex networks within a data base or between separate data bases.
- handles full text as well as structured data. As a consequence of unlimited variations in field length and frequency of field occurence, some of the system components provide a full text system not dissimilar to products like Word Cruncher (4) while some of the very components upon which this system is based, are also the backbone of a system for *nominative record linkage, the* technique for the systematic comparison of independent sets of historical sources for identical individuals occurring in both of them.
- separates clearly between data - e.g. literally transcribed terminology - and knowledge about the data - e.g. the information needed to overcome differences in spelling.
- provides specialized output functions (like cartographical ones) or supports intefacing into other packages for tasks like statistical analysis or typesetting.

### 3.2 StanFEP - A Preprocessor

While Κλειω is powerful when it comes to the handling of data structures, it expects input which usually implies, that portions of a historical source are re-arranged according to a flexible, but finite set of input conventions.

In Siegen a second independent software system, the Standard Format Exchange Program will be demonstrated in a first version, due for release in August. The user's view of this system is discussed in a separate paper contained in this folder: Thomas Werner: *Transforming Machine Readable Sources.*

This paper describes the system mentioned, as a new approach to the preparation of historical sources of the running text type for data base treatment. It shall allow the historian to transcribe a historical source without any change, loading it into data bases via structuring symbols integrated into the text.

While from a methodological point of view, we expect a significant enhancement of the quality of research possible, StanFEP, being controlled by an object oriented language, may in the context of the Siegen conferen-

ce be considered also as a tool for two separate purposes:

- the conversion of texts made machine readable for typesetting purposes into a form, which allows their analysis and further treatment by other types of software.
- the conversion of »neutral« exchange formats into the formats required by software systems processing them. A typical example for this type of application is the kind of processing required to convert data prepared according to the specifications of the Text Encoding Initiative, being supported by ALLC and ACH and presented at Siegen, into formats understood by existing software.

### 3.3 C Add-On Libraries

To facilitate the development of further software components, the decision has been taken to develop a number of C add-on libraries, which support programming techniques frequently required by the cooperating projects. This decision reflects the policy, to develop »real world« software: software, that is, which allows the handling of data bases and lexica of theoretically unlimited and practically very large size, being equivalent to systems which would consist of tens of thousand rules, usually not feasibly to be realized with current 4 GL languages.

The four libraries being either used in production systems or undergoing currently the final stages of testing are:

- Q-NET. This C add-on of I/O functions is built upon the concept of network oriented data structures. It supports at the same time
  - indexed (»keyed«) I/O similar to existing ISAM/b-tree handlers.
  - the administration of large scale dictionaries used in morphological analysis.
  - Features for approximate key comparisons are under devlopment.

In the context of the general Historical Workstation project this add-on has been used for the development of modules, which require the speedy building of interpretative systems, based on both, large specialized lexica and rule sets developed by the individual user. An introduction into this kind of module is given by the paper of Detlef v. Bovert and Rolf Huthsteiner: *Heilige Zeiten: Mittelalterliche Chronologie als Historisches Wissen*, contained in this folder. The problem they discuss deals with medieval chronology, where dates are not given on absolute time scales, but relative to variable events. (The feasts of the Saints and the variable feasts of the church.) Their paper describes the tools which are provided for the definition of special chronologies, to allow for the parsing of calender dates as closely as possible to their representation in the original source.

- CMATCH. An implementation of string oriented pattern matching functions, which allows to combine the power of SNOBOL 4's rele-

vant techniques with the speed of C-based I/O. (This add-on has been heavily drawn upon in the implementation of StanFEP introduced above.)

- FuzzNet. This add-on, which is not yet in production use, supports the building of terminological thesauri, which allow the weighting of the various relationships between terms. This add-on and the one mentioned before are prepared for interaction, to allow for the definition of »semantic« patterns, where the components making up the patterns are defined as terms related to a base term by a variable distance within a terminological thesaurus.
- TopLib. This (very small) add-on provides topological predicates, which will, after final testing, be used for both, cartographical and image based analysis of the spatial relationship between components.

### 3.4 Context Sensitive Data Bases

As part of the ongoing development work, the basic data model employed by the Κλειω software is currently being redefined. This redefinition, which is scheduled to replace the current data model by a more general in summer 1991, is introduced and discussed by Wolfgang Levermann in his paper: *Historical Data Bases and the Context Sensitive Handling of Data.* Historical information should be administered by software, which allows a clearcut separation of the character strings representing witnesses for some historical event and the varying hypotheses formed by a Historian about their meaning. This connection has to be handled:

- dynamically. Every change of the hypothesis of a researcher has to be reflected immediately by the data base in question.
- context sensitive. The evaluation of a field depends on a set of rules, which are built into the data base model, which automatically consult other field of the data base, to find all the information necessary to interpret the first one.
- fuzzy. All such evaluations have to take allow for approximate resoning.

### 3.5 Advanced Linguistic Considerations

While all the components discussed so far, are either being used in production work or in advanced stages of testing, a very ambitious long term project, funded by the Austrian ministry of science, is described in the contribution of Ursula Leiter-Koehrer: *Linguistic knowledge as a background component of an application oriented workstation*, also being included in this folder.

Unlike the other papers presented here, this one does not describe concepts which already can be demonstrated as running software in Siegen.

During the study of large scale text bases it oftenly seems, that genuine linguistic knowledge would improve the access to the data. In Humanities projects, this remains usually just a concept, however, as the implementation of tools based upon such concepts goes well beyond the resources of application oriented studies. To solve this problem a set of tools shall be provided which can be linked into existing applications and support the concept of »Semantic Parsing«. This is defined as the tokenization of running text into units which are defined by their meaning. This process is controlled by the user in a twofold process: on the one hand he/she is required to specify a formal semiotic description of the entities to be found; on the other a set of rules which binds these descriptions unto patterns which can be parsed in the text is supported.

### 3.6 Image Processing

Currently a data type »image« is being implemented within the software environment discussed here. The IBM research laboratory at Winchester has allowed the use of it's image processing library IMPART, developed there under the direction of John Ibbotson, for this purpose.

While an image handler built with the use of X WINDOWS components becomes part of the basic data base software, this library provides a large array of image enhancement and pattern recognition functions. Gerhard Jaritz describes in his paper *The Image as Historical Source,* the strategies employed by the *Institutfür Mittelalterliche Realienkunde* of the Austrian Academy of Science for the use of this software.

At this institute, since the beginning of the seventies, a largescale archive of medieval images has been collected. The 20.000 images currently in the collection have since quite some time been administered by a computer based system, allowing at the same time for classical retrieval operations as well as for quantitative and similar analysis. It is hoped, that - besides the obvious intention of immediate image retrieval - the »binding« of parts of structured descriptions of the images to their bit mapped form, will ultimately result in a more objective approach towards questions of typologies of objects of medieval life of the middle ages.

### 3.7 Manuscript Processing

Susanne Botzem and Ingo H. Kropac describe in their paper *Integrated Computer Supported Editing, Aproaches and Strategies* the concept of an editorial study, which uses from the beginning digitally scanned images of the manuscripts for transcription, uses integrated tools during their interpretation and results in a data representation which can be used at the same time for typesetting the final edition, as well as for analytical pro-

cedures. It concentrates upon the representation forms necessary to link in this way scanned data, knowledge bases and classical data base tools.

At the time of preparation of this introductory paper it is not yet clear, how far some of the techniques described can be shown at the Siegen conference.

While the project described is in it's very early stages, it is a further development of the ongoing use of computer supported technologies at the university of Graz, where since about five years computer supported technologies have been used to prepare the regional chartularty of the area for both, an improved edition as well as a prosopographical data base. The experiences gained in these projects form now the base for the further editorial study mentioned, which is realized in cooperation with the city archives of Regensburg.

## Notes

(1) Two introductory remarks on this and the following papers are in order: The collected papers following have been originally written to support huge coordinated demonstrations of a number of interrelated projects at seven European conferences between spring and summer 1990. The literature of the last 15 months has not been considered. This paper is to be understood as an introduction to the following ones; we avoid therefore to discuss the specifics of the developments at Göttingen, which in many respects form the backbone of the work described here. A selection of titles dealing with the methodological background of what is being done there, would include: Manfred Thaller »Zur Formalisierbarkeit hermeneutischen Verstehens in der Historie.« In *Mentalitäten und Lebensverhältnisse. Beispiele aus der Sozialgeschichte der Neuzeit. Rudolf Vierhaus zum 60. Geburtstag.* Göttingen: Vandenhoeck & Ruprecht, 1982, pp. 439-454; Manfred Thaller »Ungefähre Exaktheit. Theoretische Grundlagen und praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte >unscharfer< Systeme.« In *Neue Ansätze in der Geschichtswissenschaft.* Ed. H. Nagl-Docekal and F. Wimmer. Wien: VWGÖ, 1984 *(= Conceptus Studien* 1), pp. 77-100; Manfred Thaller: »Can We Afford to Use the Computer; Can We Afford Not to Use it?« In *Informatique et Prosopographie.* Ed. H. Millet. Paris: CNRS 1985, pp. 339-51; Manfred Thaller (Ed.): *Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung.* St. Katharinen: Scripta Mercaturae 1986 (= *Historisch-Sozialwissenschaftliche Forschungen 20);* Manfred Thaller: »A Draft Proposal for the Coding of Machine Readable Sources.« *Historical Social Research/ Historische Sozialfor-*

*schung,* 40 (October 1986) pp. 3-46; Manfred Thaller: »The Daily Life of the Middle Ages, Editions of Sources and Data Processing.« *Medium Aevum Quotidianum* , 10 (1987), pp. 6-29; Manfred Thaller: »Secundum Manus. Zur Datenverarbeitung mehrschichtiger Editionen.« In *Geschichte und ihre Quellen. Festschrift für Friedrich Hausmann zum 70. Geburtstag.* Ed. R. Härtel et al. Graz: Akademische Druck- u. Verlagsanstalt, 1987, pp. 629-37; Manfred Thaller »Methods and Techniques of Historical Computation.« In *History and Computing.* Ed. P. Denley and D. Hopkin. Manchester: University Press, 1987, pp. 147-56; Manfred Thaller: »Vom Beleg zum Begriff. Der Beitrag der Datenverarbeitung zur Lösung von Terminologieproblemen.« In *Ut populus ad historiam trahatur.* Ed. G. M. Dienes et al. Graz: Leykam 1988, pp. 237-54; Manfred Thaller: »Gibt es eine fachspezifische Datenverarbeitung in den historischen Wissenschaften? Quellenbanktechniken in der Geschichtswissenschaft.« In *Geschichtswissenschaft und elektronische Datenverarbeitung.* Ed. K. H. Kaufhold and J. Schneider. Wiesbaden: Steiner, 1988, pp. 45-83; Manfred Thaller: »A Draft Proposal for a Format Exchange Program.« In *Standardisation et échange des bases de données historiques. Actes de la troiséme Table Ronde internationale tenue au L.I.S.H. (C.N.R.S.)* Ed. J.-P. Genet. Paris: CNRS, 1988, pp. 329-75; Manfred Thaller: Kteuo *3.1.1 Ein Datenbanksystem* St. Katharinen: Scripta Mercaturae, 1989; Manfred Thaller: »Have Very Large Data Bases Methodological Relevance?« In *Conceptual and Numerical Analysis of Data* Ed. O. Opitz. Berlin etc.: Springer, 1989; Manfred Thaller: »Geographische Angaben in einer Historischen Datenbank.« *Eratosthene-Sphragide* 2 (1990);

(2) That Apple computers are missing from this list, may look strange to an American reader: it reflects the not so successful marketing of this manufacturer in many provinces of European Academia.

(3) Many institutions contributed to the ongoing research and devlopment work described in these papers; individual software components were contributed by a number of researchers. For acknowledgements and credits consult the individual papers.

(4) A good example of what we mean by »ugly but powerful«: these components are until version 4 controlled by fairly oldfashioned menus, are open, however, for interfaces to lemmatization systems, like one for the Latin language, developed at the Istituto Linguistica Computazionale at Pisa by Andrea Bozzi.