The Need for Standards: Data Modelling and Exchange [1991]

Thaller, Manfred

Historical Social Research
Historische Sozialforschung

Manfred Thaller:

The Need for Standards: Data Modelling and Exchange [1991]

doi: 10.12759/hsr.suppl.29.2017.203-220

Published in:

Historical Social Research Supplement 29 (2017)

Cite as:

Manfred Thaller. 2017. The Need for Standards: Data Modelling and Exchange [1991]. *Historical Social Research* Supplement 29: 203-220. doi: 10.12759/hsr.suppl.29.2017.203-220.

Historical Social Research
Historische Sozialforschung

Other articles published in this Supplement:

Manfred Thaller
Between the Chairs. An Interdisciplinary Career.
doi: 10.12759/hsr.suppl.29.2017.7-109

Manfred Thaller
Automation on Parnassus. CLIO – A Databank Oriented System for Historians [1980].
doi: 10.12759/hsr.suppl.29.2017.113-137

Manfred Thaller
Ungefähre Exaktheit. Theoretische Grundlagen und praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte ‚unscharfer' Systeme [1984].
doi: 10.12759/hsr.suppl.29.2017.138-159

Manfred Thaller
Vorüberlegungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften [1986].
doi: 10.12759/hsr.suppl.29.2017.160-177

Manfred Thaller
Entzauberungen: Die Entwicklung einer fachspezifischen historischen Datenverarbeitung in der Bundesrepublik [1990].
doi: 10.12759/hsr.suppl.29.2017.178-192

Manfred Thaller
The Need for a Theory of Historical Computing [1991].
doi: 10.12759/hsr.suppl.29.2017.193-202

Manfred Thaller
The Need for Standards: Data Modelling and Exchange [1991].
doi: 10.12759/hsr.suppl.29.2017.203-220

Manfred Thaller
Von der Mißverständlichkeit des Selbstverständlichen. Beobachtungen zur Diskussion über die Nützlichkeit formaler Verfahren in der Geschichtswissenschaft [1992].
doi: 10.12759/hsr.suppl.29.2017.221-242

Manfred Thaller
The Archive on Top of your Desk. An Introduction to Self-Documenting Image Files [1993].
doi: 10.12759/hsr.suppl.29.2017.243-259

Manfred Thaller
Historical Information Science: Is there such a Thing? New Comments on an old Idea [1993].
doi: 10.12759/hsr.suppl.29.2017.260-286

Manfred Thaller
Source Oriented Data Processing and Quantification: Distrustful Brothers [1995]
doi: 10.12759/hsr.suppl.29.2017.287-306

Manfred Thaller
From the Digitized to the Digital Library [2001].
doi: 10.12759/hsr.suppl.29.2017.307-319

Manfred Thaller
Reproduktion, Erschließung, Edition, Interpretation: Ihre Beziehungen in einer digitalen Welt [2005].
doi: 10.12759/hsr.suppl.29.2017.320-343

Manfred Thaller
The Cologne Information Model: Representing Information Persistently [2009].
doi: 10.12759/hsr.suppl.29.2017.344-356

For further information on our journal, including tables of contents, article abstracts, and our extensive online archive, please visit http://www.gesis.org/en/hsr.

The Need for Standards: Data Modelling and Exchange [1991]

Manfred Thaller[*]

Abstract: *»Über die Notwendigkeit von Standards: Datenmodellierung und Datenaustausch«*. When discussing standardization and secondary analysis in computer driven historical research, we should clearly distinguish between different approaches towards the usage of computers. The properties of four of them – statistical analysis, structured data bases, full text systems and annotation systems – are discussed and compared in some detail. Standards which want to convince users, that they should follow just one of these approaches will not succeed. What we need is a discussion of the communalities between the underlying information models and the identification of properties, for which clear conceptual models can be devised. Such clear conceptual models are a prerequisite for technical solutions, which ultimately can enable the exchange of data across the different approaches. Such conceptual models, therefore, are what we need as standards.

Keywords: Standardization, information modelling, data bases.

1. Introduction[1]

If in 2086 the Association for History and Computing commissions a volume to celebrate its first hundred years, the appointed author will undoubtedly begin research on "Part I: The Formative Years" by plundering the funding proposals written by historians from about the late-1950s. There will be discovered with astonishing regularity the same bold claim: "this project needs special attention because it is important in its own right and because it will make available a huge corpus of historical material of central importance to the discipline". A contradictory theme will emerge when the same author gets round to a thorough review of the conference proceedings and journals prepared by and for computer-literate scholars in history and other humanities discipline say before 1991. For there will be found the persistent lament that scholars make too little use of that very machine-readable infor-

[*] Reprint of: Manfred Thaller. 1991. The Need for Standards: Data Modelling and Exchange. In: *Modelling Historical Data* (= Halbgraue Reihe zur Historischen Fachinformatik A 11), ed. Daniel Greenstein, 1-18. St. Katharinen: Scripta Mercaturae.

[1] This paper has been submittted to the editor in the author's usual Germlish. In turning it into proper English, the editor of this volume has considerably improved it, not only stylistically, but also in substance.

DOI: 10.12759/hsr.suppl.29.2017.203-220

mation whose compilation was justified in part with the promise of secondary analysis.[2] To scholars working with computers in the 1990s, the contradiction is easily explained. Analyzing someone else's dataset is difficult. In fact, it is so difficult that it is often simpler and less costly to create one's own dataset from scratch rather than use one prepared by another researcher from the same underlying documentary sources. This unnecessarily expensive and somewhat irrational behaviour has not gone unnoticed by computer-literate historians or by agents of funding bodies. At least both groups have paid lip-service to the idea that some sort of standard for encoding machine-readable information is a good thing, and several independent initiatives have recently been launched to define such standards. Strangely, however, such initiatives have yet to bear fruit. In this paper, I will argue that attempts to define suitable standards have so far failed to produce practical results because they have concentrated narrowly on defining simple input formats or tag sets[3].

Computers promise the historian the means of integrating into a scholarly investigation more information than was hitherto thought possible. Already they have opened up whole new areas of research and illuminated new aspects of the past by enabling historians to exploit sources whose bulk and complexity made them inpenetrable to manual analysis. In future, it is likely that the benefits will be extended when each of us can have access to datasets compiled by colleagues conducting research in related areas or with related documentary sources. Census-based urban studies, for example, will benefit from access to machine-readable data compiled for similar studies of neighbouring townships or simply by using the occupational coding schemes already implemented in earlier studies and openly available through online databases. The computer-literate psephologist will log onto online databases to extract voting data relevant to a particular investigation, supplementing them only with those data which are necessary to the study but not yet available in machine-readable form. The historian struggling to interpret the cryptic abbreviations which are so common in medieval documents will get help from machine-readable data dictionaries established to decipher the catalogue of abbreviations found in a computer-assisted study of related sources. But this promise can only be kept if machine-readable "information" can be exchanged easily between individuals who may have radically different research aims historical sources frequently contain information which may be as profitably

[2] Concern was already being voiced back in 1962: cf. Jean Claude Gardin, "The use of Computers in Anthropology" in Dell Hymes (ed.), *The Use of Computers in Anthropology,* London 1965, 99.

[3] During the past 15 years the author has been connected with the development of Κλειω – software which is specifically tailored to the needs of historians. The purpose of *this* paper, however, is neither to explain nor promote how Κλειω approaches the problems of data modelling and exchange but to show how software which uses rather different approaches to the same problems might fruitfully coexist. At the same time we cannot very well conceal the fact that Κλειω and the *Historical Workstation Project* attempt to implement in a practical way the rather more abstract and theoretical propositions about data modelling and exchange outlined in these pages.

exploited by statistical analysis as by text searches – and that means ultimately between different computer processing environments which insist upon differently structured data. In this respect, the future relies in part on "standardization".

But it is not sufficient merely to define standard conventions for structuring machine-readable information. For if we want to create an infrastructure which allows historians to draw upon a vast array of machine-readable data, we will also have to convince the individual researcher to adopt whatever standards are devised so that the fruits of every computer-aided investigations can be contributed to the ever-growing pool of machine-readable resources. The problem here is that at present professional historians are pressed to produce results. Those that use computers in there research will only accept standards whose implementation doesn't prolong the research process or constrain the development and use of specialist techniques. The political historian interested in determining the composition and inter-connectivity of ruling groups from autobiographies is hardly likely to abandon the tried and tested method of selecting a finite amount of data from the autobiographies and entering them into a tabular format appropriate to databases and statistical packages. At the same time, the political historian interested in the same autobiographies because they illuminate the political outlook of ruling groups is hardly likely to enter the text of the autobiographies into a statistical package or relational database in order to conduct concordance and collocate analyses. Moreover, since it is rarely possible from the outset of a research project to predict precisely what "relevant" information will be encountered in the documentary evidence, and how that information should be structured and later analyzed, no scholar will feel comfortable with a standard which seems to infringe upon his or her liberty to create solutions for situations which have yet to be encountered. A standard should not, in other words, *prescribe* to scholars how they should solve a given set of problems. The approach to standardization proposed by the Text Encoding Initiative (TEI) and developed elsewhere in this volume is I feel, *prescriptive.* It offers stock solutions to a finite set of problems associated with computer-aided data management and analysis and instructs the researcher to use these solutions regardless of the nature of his or her sources and analytical aims. It is also my view that the user of *prescriptive* standards like the consumer of ready-made clothes sacrifices a good fit for the sake of convenience.

The alternative approach to standardization which we develop in this paper does not launch out from a finite set of model solutions and attempt to adapt these however uncomfortably to different data management and processing problems. It is rather more consumer-oriented and based on our belief that historians, like the political historians mentioned above, structure machine-readable data differently according to their very different analytical aims. A standard must accommodate each of these very different aims, but at the same time facilitate automatic conversion of data structured for very different uses. From this perspective, standardization needs to be *descriptive* rather than *prescriptive.* Its proper domain is not prescribing to scholars how they should solve a given problem but furnishing scholars with the tools necessary to describe in an unambiguous way the problems they encountered when creating their datasets and how they chose to solve them.

When such a standard is available, only then will it be possible to create the converconversion facilities which will allow interchange of data modelled for use with very different processing environments.[4]

There are currently at least four radically different processing environments which historians use and which rely upon differently structured data. Statistical packages on the whole rely upon what I shall refer to as an interpretive data structure where numerically pre-coded data are presented in a tabular format. Databases rely on structured data where the researcher doesn't pre-code information but is highly selective in what information is taken from the source and entered into database table(s). Full-text systems record the entire text of the historical source(s) in question and use a comprehensive tagging scheme to indicate where categories of information begin and end. This data structure I refer to as pre-edited. The fourth processing environment is that associated with so-called hypermedia systems and relies on an "image-bound" data structure. By outlining the central properties of the data structures appropriate to each of these processing environments it is possible to determine which properties the data structures share in common and to recommend how conversion between them – the central aim of a consumer-oriented standard – may be made possible. The essential properties of the data structures outlined above are perhaps best described by showing how each would organize information drawn from the same historical source. For this purpose, I have chosen a hypothetical report about monetary obligations connected to land holdings which the historian would want to administer in a database. A fragment of the source might read:

> At Michaelmas in the fortieth year of Elizabeth, our gracious Queen, Johnny Turner, who holds the smitthy behind the church, has paid the seven guineas he owed to his brother Frederick from the will of his late father, and all his property is now free of any obligations whatsoever.

In a statistical package the data might look like this:

```
09071598 John               Turner                              456 0147 110 7 0
Month:                      columns:       01        - 02
Day:                        columns:       03        - 04
Year:                       columns:       05        - 08
Christian Name:             columns:       10        - 29
Surname:                    columns:       30        - 49
Occupation:                 columns:       50        - 52
Amount in sh.:              columns:       54        - 57
Payment to:                 columns:       59        - 61
(1xx == Relative
11x == Brother)
Type of obligation:         columns:       63        - 63
Obligation remaining?       columns:       65        - 65
```

This data structure is very well known but probably not so widely used as it once was. It is easily identified by the numerical codes which are used to represent

[4] The preference for descriptive versus prescriptive standards is developed with respect to well-known domain of occupational coding by Daniel Greenstein, "Standard, Meta-Standard: A Framework for Coding Occupational Data", *Historical Social Research* 16 (1991) 1, 3-22. doi: 10.12759/hsr.16.1991.1.3-22.

information whose meaning needs to be interpreted by the historian before it can be coded and entered into the database. Let us quickly summarize the properties of this *interpretive* data structure:

- The physical characteristics of the source are ignored.
- The syntax by which relationships between individual chunks of information are expressed in natural language is ignored.
- Translating the semantics of the source into more precise categories depends upon the researcher's subjective "understanding" of the source.
- Formalized treatment of the source, e.g. by means of statistical analysis, is extremely easy: the computer is seen as a tool to analyze relationships between clearly defined categories of information.
- Exploratory treatment of the source, finding out what an ambiguous term *might* mean, is well nigh impossible – once you have decided what a term means it is hard to find out that you have been wrong.
- Exchanging data between similarly interested research projects is extremely easy: FORTRAN formats are software independent and where common numerical coding schemes are employed it is unimportant from what sources or research projects data are derived.
- Exchanging data between differently interested research projects is sensible only in rare cases. A machine-readable version of tax lists compiled to analyze the development of social classes will be useless to anyone interested in the tax lists to illuminate the movement of capital between economic sectors – numerically coded categories of information will have been devised along different and incompatible lines.
- Projects which employ this method of data processing are usually completed on time. With the application of a little common sense the researcher can code and enter data quickly and estimate very accurately the amount of time that data entry is likely to take.
- Data entry is error prone due to mistyping and miscoding, and it is hard to find errors once they have been made. Datasets may therefore be unreliable.

There is no need here to rehearse the well-developed debate about the strengths and weaknesses associated with this kind of data processing.[5] Most historians today would avoid such an environment and opt instead for some kind of database management system. While quite a few database systems are available, let's take as an example one which relies upon a more or less self-explanatory "descriptor – descriptum" or "tag – content" logic. We introduce into it an additional convention which appends as a suffix to interpreted terms the original spelling in appropriate cases. The interpreted and original terms are only separated by an at-sign (@). The data shown below are divided into three hierarchically- related "entities": "payment", "holder", and "addressee".

[5] See Mark Overton: "Computer Analysis of an Inconsistent Data Source: the Case of Probate Inventories", *Journal of Historical Geography* 3 (1977), 317-26.

```
payment$date=7 Sep 1598@Michaelmas 40 Eliz I/sum=7g/
      furtherobligation=none/reason=will
holder$surname=Turner/cname=Johnny/occupation=blacksmith@
       holds the smitthy behind the church
addressee$cname=Frederick/relationship=brother
```

In this *structured* environment selected portions of information are taken from the original source and entered into the appropriate database "field" more or less unchanged. The main characteristics of this data structure are:

- The physical characteristics of the source are ignored.
- The syntax by which relationships between individual chunks of information are expressed in natural language is ignored.
- The semantics of the source are translated into more precise categories with the aid of various software tools: machine-readable code books or look-up tables which translate terminology found in database fields into numerical codes; algorithms which normalize variant forms of, for example, chronological references; a combination of algorithms and look-up tables for normalizing categories of information, for example, currency references, before they are numerically coded, etc.
- Formalized treatment of the source, e.g. by means of statistical analysis, is possible but will depend almost entirely on the extent to which the software permits the construction of the look-up tables and algorithms mentioned above. Generally, however, such databases are used as tools which help to derive formal categories and to prepare them for analysis often in an interpretive processing environment (statistical package).
- Exploratory treatment of the source, finding out about what an ambiguous term *might* mean, is extremely easy and is one of the principal reasons that historians will use this processing environment.
- The production of meaningful lists and catalogues from the categories of information stored as fields in the database is also quite simple. But where such lists comprise several and / or very lengthy fields it may not be so simple to print them out neatly onto a page. Text-formatting facilities are normally extremely limited.
- Exchanging data between similarly interested research projects is possible. As a rule, what goes into one data base can be written out as an ASCII file and read into another[6]. However, where two projects have made slightly different decisions about what it means to, for example, "transcribe information *almost* as it appears in the original", data exchange can require an astonishing amount of sophisticated programming.[7]
- Exchanging data between differently interested research projects is no more difficult than between similarly interested ones, but it is not easier either.

[6] Cf. in a wider context also Leen Breure, "Historische databasesystemen", in Onno Boonstra et al.: *Historische Informatiekunde*, Hilversum: Verloren 1990, 123.

[7] Cf. Rainer Metz: "Global Data Banks: a Wise Choice or Foolish Mistake", *Historical Social Research* 16 (1991) 1, 98-102. doi: 10.12759/hsr.16.1991.1.98-102.

- There is always the danger that an unreasonably large amount of information will be recorded in a maze of machine-readable database tables. Projects therefore require far more attention to careful project management and pilot study than those which use interpretive data structures.
- The reliability of the data is very high which is why this processing environment was invented in the 1970s in the first place.

Full text processing environments require a *pre-edited* data structure which incidentally, is the data structure in which SGML is most commonly associated and which provides the main focus of the TEI. The same pre-edited approach was employed extensively in the 1970s by Alan Macfarlane in the "Earls Colne Project" – an attempt to create a machine- readable transcript of the surviving pre-1870 documents from one English village. With each document Macfarlane hoped to transcribe the whole text, adding additional markup to delimit specific categories of information.[8] Unfortunately Macfarlane's enormous undertaking did not produce easily-accessible documentation, thus rendering the surviving database more or less unapproachable. The software supporting the project was way ahead of its time and the tagging conventions actually incorporated most of those now supported by SGML, plus a few (notably those dealing with backward references) which are at best very difficult to express in SGML. Entered up in the *pre-edited* form our example data might look as follows:

> (payment (date At (feast Michaelmas feast) in the (year fortieth year) year of (ruler Elizabeth ruler) date), our gracious Queen, (holder (cname Johnny cname) (surname Turner surname), who (status holds the smitthy behind the church status) person), has paid the (amount seven guineas amount) he owed to his (addressee (relationship brother relationship) (cname Frederick cname) addressee) from the (reason will of his late father reason), and all his property is now (further free of any obligations further) whatsoever. payment)

The properties of this data structure may be characterized as follows:

- The physical characteristics of the source are ignored.
- The syntax by which relationships between individual chunks of information are expressed in natural language is explicitly used to interpret that information. The amount of explicit markup required in addition to that already present in the form of normal writing conventions depends on the quality of the parsing software being used. The practicality of this approach is also largely dependent on the quality of the parsing software and on the availability of tools which aid in the insertion of complex markup schemes.
- The semantics of the source are translated into more precise categories with machine- readable codebooks and algorithms not unlike those used by structured processing environments.

[8] Alan Macfarlane: *Reconstructing Historical Communities,* Cambridge 1977; Donald E. Ginter et al., "A Review of Optimal Input Methods: Fixed Field, Free Field and the Edited Text", *Historical Methods Newsletter* 10 (1977), 166-76; Timothy J. King, "The Use of Computers for Storing Records in Historical Research", *Historical Methods* 14 (1981), 59-64.

- Formalized treatment of the source e.g., by means of statistical analysis, is possible but contingent on the quality of the codebooks and algorithms mentioned above. Generally the *pre-edited* data structure is best suited to textual analyses.
- Exploratory treatment of the source, finding out about what an ambiguous term *might* mean, is extremely easy.
- Formatting the machine-readable text or any part thereof for printed production is also quite simple, and this processing environment is frequently used by projects for which printed output is of primary importance. But as many researchers have discovered, transforming a dataset which is appropriate for printed production and textual analysis to one appropriate to more formalized statistical analysis requires more work after data entry has been completed than many are prepared to accept.
- Exchanging data between projects is possible but requires sophisticated software. Predicting the compatibility of the tagging conventions used in different datasets can be an extremely tricky business. However, with the pre-edited environment a project's research aims have no bearing on the possibilities for data exchange since in any event, datasets comprise complete machine-readable transcriptions of their underlying sources.
- Researchers who use the pre-edited data structure often have a very clear picture of how their printed output should look but are less clear in their analytical aims.

Because decisions about analytical aims have far-reaching ramifications for how text is tagged, project management and pilot investigation take on the greatest importance.

While the strengths and weaknesses of the three data structures outlined above are fairly well understood those associated with the *image bound* data structures used in HyperCard, HyperMedia, HyperTalk ... in other words, the Hyper World, are far less so. Unfortunately, an illustration of how our sample data when prepared for Hyper use is not easily provided in print and can only be described in the following general terms:

a) The document itself is contained in the data processing system in scanned, that is bitmapped, form.

b) The categories of information delimited for analytical purposes in the other data structures discussed above could be represented and a link established between the delimited categories and the appropriate section of the bitmapped image from which they were derived.

c) It is possible to move from any of the delimited categories (attributes, fields, or whatever your preferred terminology) to the original graphical representation of that part of the source from which the categories were derived and to do this right on the same screen.

d) It is also possible in theory (though much rarer in practice) to look at a portion of the scanned document and go from there to the interpretation of that specific portion as represented in the delimited category (attributes, fields, ...) which was derived from it.

Ignoring the messianic statements that the Hyper crowd, like the proponents of any new technological innovation, are prone to make, I would propose the following as the central properties of the *bound image* data structure.

- The physical characteristics of the source are maintained.
- The syntax and semantics of the sources are maintained but in varying degrees. Indeed, bound image systems may simply situate a special data type (one called image) on top of structured or pre-edited data. The resultant dataset spectacularly changes the data's visual appearance but adds few other processing capabilities.
- Formalized treatment of the source e.g., by means of statistical analysis, is usually not central to research projects which use a bound-image data structure. Indeed the visual spectacle of the bound-image environment often obscures its analytical potential from the researcher's view. Hopefully this is a transient phenomenon. Logically, the tools for formalizing information that is bound to an image are no more difficult to realize than those associated with data structures which haven't any such links.
- Exploratory treatment of the source, browsing through the source (frequently up to the point of getting lost in it), is what such systems are principally used for.
- Data exchange between projects, whatever their research interests, takes on an entirely different set of problems as those hitherto encountered. Exchanging images is especially difficult because it may be dependent on the hardware *and* the software used by the different projects. The main problem here, however, seems to be, that the necessary expertise is relatively rare. In principle standards for image files – like TIFF – are much more oriented toward interchange than standards for character representation were at a more mature stage in their development. More problematical is the fact that there seems to be virtually no notion that image-bound systems may have some analytical interests. Consequently, techniques for refining HyperCard stacks into statistical cases remain undeveloped.

These four data structures and their associated processing environments exist with equal right. It would be extremely hard to convince an econometrician to prepare an early-twentieth-century list of production statistics to use the pre-edited environment exploited by Macfarlane. On the other hand, if you have just spent a few million dollars to convert into a bound image dataset an entire archival collection of documents relating to early Spanish colonial expansion, you might manage to remain civil when told that standardization entails describing the categories you used in a specific "variable" of your data set.[9]

[9] The reference here is to a Spanish project to convert the entire contents of the Archivos de las Indias into a bound-image type database. While no recent published descriptions of the project are known to the author, information from Spain suggests that the material will almost certainly be scanned by 1992, in time for the anniversary celebrations of Columbus's discovery of the Americas. Description of the material, the process of actually binding images to structured interpretations of them, will, however, continue for quite a few years yet.

Now, admittedly, relationships between the letters of Christofero Columbus and Polish steel production in the 1930s are few and far between. But relationships between the kinds of data being structured differently are all but artificial. Those between the mandates of the Castilian crown and the trade between Sevilla and the Americas in the 16th and the first half of the 17th century are much more apparent. And, while the Chaunus published their tables as bound volumes[10] in the 1950s, their quality being just below the level of OCR today, these volumes might quite easily be read after the next advance of OCR technology; having sufficient structural markup in the printed layout to facilitate parsing them into a database without the addition of much explicit markup. Indeed, while most bound image projects are currently dedicated to browsing through data rather than to analyzing it statistically, some have the potential to express in numerical terms things like the relative placement of various motifs on images thereby opening up whole new possibilities of formalized analysis. And while we have argued the extremes to emphasize the ultimate unity of the field, the possibilities of mutual reference between datasets prepared in a structured way and those prepared by pre-editing markup are probably much more obvious than the ones between interpretive numerical datasets and bound images.

All this leads me to suggest that there are two kinds of standards, both of which are necessary. For the adherent of a specific type of computer usage, standardization will entail the development of generally agreed rules for that type of computer usage. This approach to standardization might produce rules governing how machine-readable information should be structured in statistical, relational or full-text databases. Such standards would be extremely appealing to the technical experts who want to define how to proceed in their own specialist areas. But when you look at the technical landscape as a “consumer” you might decide that each of the processing environments may at one time or other be appropriate. This is precisely the case for the consumer who also happens to be an historian. What the polymath consumer wants from a standard is not, or not only, some instruction about how to become agreeable to one of the main communities of computer users. Rather, the polymath interested in our example data, for instance, would want to (a) find out whether blacksmiths paid off their inherited obligations more quickly than other occupational groups, (b) get from court records an alphabetical list (orthographical fancy and all) of all people living in a particular village, (c) analyze whether any change occurred in the literary context surrounding references to royalty,[11] (d) to place a sensible caption under a scanned drawing of a village smith’s shop which will be the opening screen in an interactive teaching unit, and

[10] Pierre et Huguette Chaunu, *Seville et l'Atlantique*, vols. 1-8 (Paris, 1955-1960).

[11] It is significant that textual analyses have recently received the most methodological attention. Virtually the only recent methodological contribution to the *Journal of Interdisciplinary History* has been the paper by Mark Olsen and Louis-Georges Harvey, "Computers in Intellectual History: Lexical Statistics and the Analysis of Political Discourse" 18 (1988), 449-64.

(e) do whatever seems appropriate with the computer currently being invented in some laboratory or garage when that computer becomes available.[12]

There is as yet another reason why standards should attempt to unify ongoing work in different processing environments by facilitating data exchange between them rather than to emphasize the differences between user communities engaged with distinctive computer methods. In the corpus of methodological articles published by computer-using historians since the late 1960s it is difficult to distinguish very much in the way of consolidated methodological advance. This is probably because many such articles fail to generalize from descriptive accounts of "how I used software LMN during my summer vacation" to address the wider intellectual and methodological implications of their work. By the early 1980s the field of history and computing was consequently littered with interesting accounts of different projects but still hadn't any rigorous intellectual or theoretical underpinnings. Lacking any definitional boundaries or intellectual rigour, the field of history and computing was flooded after the introduction of the microcomputers with a community of computer users who simply rediscovered for themselves the principals and problems already documented by the late 1960s but never formalized into an accessible, instructive and coherent literature. The effects were devastating. No recent number of *Historical Methods*, the pioneering journal which contained the most important papers on computer usage in the 1970s and early 1980s, contained any discussion of the wider implications of HyperText and other nonlinear systems. The journal has instead more or less abandoned discussion of computational techniques for statistical methods and other interdisciplinary matters. Where very occasionally computational issues are addressed, they are addressed in a manner which assumes that readers are computer novices. These developments have not happened by chance. Rather the editorial made them explicit in a 1985 number of the journal. There Daniel Scott Smith stated that the introduction of microcomputers and of undergraduate teaching in "computer literacy" would be such important events, that *Historical Methods* would actually modify its editorial policy and nominate a special editor (Janice L. Reiff) exclusively responsible for those subjects. By 1987 the journal's new emphasis was baldly apparent in two papers on text processing and computer-aided instruction.[13] Both papers undoubtedly have their value but only further detract from any approach to consolidate or advance methodological discussion. The journal's "emphasis on microcomputers and teaching" culminated in a 1988 special issue entitled *Special Issue: History, Microcomputers, and Teaching* which, after two years still could not say very much more about this new and important development than: "The articles that follow provide a variety of examples of ways in which microcomputers have been used to do such teaching. It is hoped that future issues of *Historical Methods* will contain articles that demonstrate useful and imaginative ways to use these

[12] It is extremely important to emphasize that the experience of the last two decades has taught us that any "standard" which is tied to a specific level of technological development is bound to fail.

[13] Richard Jensen, "The Hand Writing on the Screen", *HM* 20 (1987), 35-45; Richard C. Rohrs, "Sources and Strategies for Computer Aided Instruction", *HM* 20 (1987), 79-83.

machines in both teaching and research." These resounding words were followed by the almost complete disappearance of computer related articles from *Historical Methods*. We describe this phenomenon at length to emphasize that when two ways of using a computer are not understood as being very closely related parts of a larger field, they have a tendency to grow apart as very distinctive sub-disciplines and to fragment the discipline. "Standards" for dataprocessing in history, which do not strive for cross-fertilization between the "subcultures" of computer usage – whether these are defined by their orientation toward teaching or research, or by their respective processing environments – "standards", which are directed at a specific part of the wider user community will be harmful, not productive.

So this second "consumer-oriented" standard is not only a prerequisite for data exchange but essential if the unity of the discipline is to be preserved. But is it practicable or just a utopian pipe dream? The difficulties in developing a consumer-oriented standard cannot be underestimated. Not the least of these is that technical experts engaged with one type of data processing tend to think of all other types simply as degenerated cases which are in principle already catered for in their own favoured environments. One assumes, however, that while narrow technical standards may be easier to define, the second variety is the only one which will ever be accepted by researchers themselves. A second obstacle is that a utopian consumer standard need to be built upon a *conceptual* understanding of the kinds of data found in and used by specialists within a given academic discipline, in this case, in history. The specialist's *prescriptive* standard is derived from a conceptual understanding of specific computer related techniques. The problem is that whereas the conceptual understanding of specific computer models is already well developed, historians haven't yet developed a formal understanding of the datatypes that they so frequently encounter in their computer-aided research. To flesh out this abstract distinction between the two different kinds of standards, we will look more closely at what happens when you try to turn the contents of statistical "variables" into database "fields" and then into tagged information as might be found in a pre-edited full text system.

In the case of statistical datasets containing interpretive numeric codes, variables are atomic. That is, between the variable referencing the day, month and year of a date, for example, there is just the same relationship as that which exists between the year of birth and occupation of the person described in any one record. So if you want to copy a "date", you would have to formulate three independent copy operations, e.g. COMPUTE commands in SPSS. In most systems using structured data – typically data base systems – you would use a specific data type "temporal information". So by using a "date" with any command for copying or any other manipulation, you would always use one command to address the three logical components as one entity even though a date still has a day, a month and a year component. Practically this could mean, that, going from a statistical package to a data base, you would have to write an output procedure which turns the content of three variables into a string, the parts of which are delineated by a separating character and vice versa. Similarly in most data base systems, "note" or "comment" fields are just atomic fields which have no implicit technical relationship to any other field, except that it might happen to be one of several fields which characterize a particular entity. For example, assume the itemized records in a tax list show only the amount of taxes due for those individuals who have paid up in full, and that in a few cases there is a scribbled note which indicates that only a specific part of the total amount had been paid. In most data base systems the two amounts of money would be represented in two fields (amount owed and amount paid, respectively). The relationship between these two fields would be no different than the relationship between the amount paid and, for example, the Christian name of the taxpayer. In systems using pre-edited markup, however, you could simply put both amounts of money into one bracket, indicating that both amounts form the "tax", being subdivided further into "amount owed" and "amount payed" thereby implying that when copying information on "tax" that both amounts need to be taken care of. We shall not continue with bound images in this respect: though they obviously rely upon additional relationships between individual chunks of data.

What should be clear by these examples is, *why* it is so difficult to devise standards which do not explicitly take into account the fact that very different processing environments require very differently structured data. For the statistical package the notion of a datatype is almost meaningless – or at least trivial. For a structured database package which can handle the temporal information "Michaelmas 40 Eliz I" or something similar, documenting the precise rules used to normalizing such dates for analytical purposes is positively central to exchange. But where does that bring us? When statistical software does not have a data type "temporal information", it does not. So how are we to exchange data between systems which "know" about temporal information, for example, and those which do not? With such profoundly different processing environments does the whole notion of exchange rapidly recede into the realm of fantasy? In our opinion this is certainly not the case. A proposal for a *descriptive* standard which starts from the problems that the consumer wants to tackle and not from a specific technical solution being propagated would focus on the problem of temporal information as follows:[14]

1) There exists the phenomenon "temporal information". Temporal information can (in a simplified way) be described by an arbitrary subset of the following conceptual variables:

- Day of the month.
- Reference point within a notational system (e.g. "Michaelmas").
- Systematic offset within a calendar system (e.g. "Tuesday after x").
- Absolute offset from a reference point (e.g. "the day before").
- Era within a calendar system (e.g. the reign of Elizabeth the first).
- Name of the month.
- Year within a calendar.
- Applicable calendar (e.g. the Islamic one).
- Offset within an era (e.g. year of reign).

[14] See the more detailed proposal by the author in "A Draft Proposal for a Standard for the Coding of Machine-Readable Sources", *Historical Social Research* 11 (1986) 4, 3-46, doi: 10.12759/hsr.11.1986.4.3-46 and elsewhere in this volume.

- Symbolic macro formulated as an expression out of the preceding terms (e.g. “dedication day” to be expanded to “Sunday after Ascension”).

2) There exists a finite number of mappings between possible subsets of these elements. These mappings can be described by:

- A number of algorithms, each converting one representation into another.
- A number of “lexica” or “look-up tables”, describing the meaning of a set of instances of one of these subsets in terms of other elements.

Okay, SPSS will still not understand *Michaelmas 40 Eliz I.* Or will it? In principle, if we take the preceding paragraphs seriously there is no reason, why SPSS could not understand the expression *Michaelmas 40 Eliz I* so long as there was a complete lexicon of all feasts on a local calendar and one giving the years during which different rulers reigned. With such lexica we could mechanically map various terms into nominally scaled variables thereby converting *Michaelmas 40 Eliz I* into something like *1245 40 126.* And, while this author would not necessarily like to do it, one *could* implement the necessary conversion algorithm as a huge set of COMPUTEs, RECODEs, IFs and the like, *provided a well understood algorithm existed.* A generic description of this approach to standardization might then be described as follows:

A “standard” for data collection and exchange, should define a set of basic “data types” of the kind just discussed. It should further identify various types of software systems – statistical packages, database administrators, natural language parsing systems, administrators of nonlinear data – which have clearly defined capabilities for treating each of the data types that are conceptually defined.[15] Now, such a “standard” does not prescribe for researchers how to organize and enter information in machine-readable form, but it does allow them to tell a fellow researcher interested in using the dataset precisely how it was constructed and the decisions that were made when data were collected, organized and entered into the system. That kind of standard must not simply consist of paper documentation but of software which can convert the representation of like datatypes (e.g. temporal information) into a format suitable to any number of different processing environments.[16] Recent discussions have brought to light seven data types which if developed conceptually would be sufficient to describe the entire range of information that historian’s encounter. They are:

- Full text.

[15] This paper intentionally avoids making clearcut distinctions between types of software systems. In practice the only important distinctions are those which distinguish the data structures used by a given system. A good example of how elusive more precise categorization of software systems can be is found in the discussion of text bases by John J. Hughes, *Bits, Bytes & Biblical Studies*, Grand rapids 1987, 497.

[16] Format conversion tools might be provided as frontends to data base software. Cf. Martin Gierl, Thomas Grotum and Thomas Werner, *Der Schritt von der Quelle zur historischen Datenbank. StanFEP: Ein Arbeitsbuch,* St. Katharinen: Scripta Mercaturae 1990 (= *Halbgraue Reihe zur Historischen Fachinformatik A* 6); Kathrin Homann: *StanFEP. Ein Programm zur freien Konvertierung von Daten,* St. Katharinen: Scripta Mercaturae 1990 (= *Halbgraue Reihe zur Historischen Fachinformatik B* 6).

- “Descriptors” or fields coded in a “controlled vocabulary”. This may seem the most obscure category for the statistical minded since it includes information as diverse as occupations, educational careers, qualifications, degrees and the like. The point is, that while in statistical systems such terms map onto some system of numeric values at the nominal level of scaling, in other worlds of software – e.g. within databases – they are frequently treated by complex networks of interrelated thesauruses or look-up tables and are thus structurally very similar.
- Names. Once again these are simply instances of nominal variables in statistical systems. Outside of this processing world, however, they are usually represented by strings between which algorithmically expressed degrees of proximity exist.
- Numbers. These are actually much more complicated in historical sources than calendar dates: being non-decimal, expressed in layered systems of units and so on.
- Calendar dates / temporal units.
- Spatial references.
- Images.

A *descriptive* standard should define conceptual models for each of these abstract data types which would then inform how information is structured within a specific processing environment in accordance with the data modelling capabilities of that environment. A specific implementation would also include a defined set of importing and exporting procedures with which data could be converted into neutral formats or formats which are more congenial to other processing environments.

Our original idea has been to create a standard not by prescribing to historians how to behave, but by describing their behaviour and their use of several different processing environments each of which relies upon radically different data structures. From this perspective, the only generic solution to standardization seems to be one which attempts some abstract definition of the types of information which actually occur in historical sources and which are represented rather differently in different processing environments. These conceptual models then should underlie all practical recommendations about how data should be organized and entered into machine-readable form in different processing environments. They are also essential to the development of conversion facilities which will enable data organized for one processing environment to be translated into a format suitable for other rather different ones. *The descriptive, consumer-oriented standard is therefore also a model driven* one. But does its promise justify the enormous effort which will be required simply to define the conceptual models let alone to create the conversion facilities? We should remind the reader what is at stake. Without the conceptual models of different historical data types the aims of standardization (any notion of standardization) cannot be achieved. Recap these briefly: secondary analysis; more fruitful cooperation between colleagues at different universities; the ability to change processing environments in mid-project; the ability to continue ten or fifteen years from now work begun today on soon-to-be outmoded hardware and software. There is no question *whether* we have to undertake the effort; the question is, how to organize it. And here we might derive some valuable lessons from previous unsuccessful attempts at defining standards.

Two reasons seem to lay behind past failures and both can be described as vicious circles. Firstly, there is the problem of secondary analysis. So long as secondary analysis remains so difficult as to be almost unapproachable, few people will attempt it. Similarly, so long as only a few people actually use a standard which includes written documentation and a supported environment, there is no reason to produce one. And so long as there aren't any software tools necessary for effective data exchange and thus for secondary analysis, secondary analysis will continue to be difficult and unapproachable and few will attempt it. Secondly, there is the problem of database preparation. So long as databases are not proofread just as painstakingly as printed monographs, they will not be trusted and used very much for secondary analysis. As long as databases are not trusted and used, their creation and more importantly their refinement through painstaking editing will provide only the smallest bonus for the academic reputation of an historian. As long as the creation of databases offers few benefits in terms of one's career, databases will not be made as free of errors as printed editions are and so will neither be trusted nor used for secondary analysis.

Work on the consumer-oriented standard outlined above needs to be undertaken in three concurrent phases.

- We need well understood and rigorously documented conceptual models of the historical datatypes defined above.
- We need tools to transfer data from one processing environment to another. No historian will follow any kind of standard which offers no guarantee about enhanced prospects for data exchange.
- We have to teach users why they might profit from the standards that are developed – e.g. how they might benefit from secondary analysis. This is not a problem amongst like-minded members of small user communities. One reason for SGML's success is its immediate appeal to publishers who have a clearly defined purpose in mind: printing books in different editions and different layouts, without reentering the text. But with the wider community of computer-using historians such instruction is central if a standard is to be developed *and* widely used. As long as secondary analysis, for example, is just a vague idea, people will pay lip service to it but certainly not invest their time, effort and limited finances into actually trying it. And without training courses which show researchers how they can get results out of somebody else's data, it will remain hard to convince anybody to prepare their own data for such purposes.

We started by saying that a standard must not be prescriptive. Computer applications *are* a growth area in historical research and a standard must never prevent somebody entering the computer-using community right now and inventing a new and better solution. The reason, why this is usually not emphasized as much as it should be is that researchers discussing standardization frequently have "guru status" within their respective communities. And, when you are asked six times a month "Very interesting, but just precisely *what* shall I *really* do in this specific situation?" you tend to forget that the "obvious answer" that one formulates on the basis of fifteen or twenty years of experience in a specific subsection of the field might be all but obvious to someone who entered data processing from a different angle. In other words, for reasons of methodological purity we should abstain from giving recipes. But all of us are asked for recipes some of the time. So what we need are two completely different types of documents and approaches to standardization. On the one hand we need standards which describe to someone expert in a particular processing environment, for example statistically oriented types of processing, what to do with information which comes in a format specific to an altogether different processing environment, for example, a HyperCard stack. That kind of standard can only be built upon a careful analysis of the common core shared by otherwise distinctive processing environments and that core we have identified as the fairly atomic conceptual models of specific types of historical information. But standards available at this level are alone insufficient. For a large number of "customers", as opposed to technical experts, a strange situation exists. They are at once interested in expressing their problems in specifically historical terms but not in abstractly conceptual ones. Thus the abstract conceptualization of historical datatypes will be too technical and of little interest to the user who simply wants to know that defined connections exist between whatever processing environment he or she chooses and the other processing environments used by the rest of the computer-using historical community. To resolve this contradiction, I think that it is the responsibility of experts in the various branches of data processing to develop standards at the level of model solutions to generic problems – solutions which are based on their knowledge of exchange possibilities between different processing environments. Such model solutions will, for example, give some guidance to users who are interested in analyzing data from a parish register using software XYZ, or compiling a database of wills with database ABC, and so on.[17]. Such model solutions would *not* express inflexible opinions about how to analyze a specific source but give some guidance by demonstrating the current state of the art, and would almost surely be used by quite a few newcomers to computer-aided historical analysis.

Computer usage in history and probably in most of the humanities, is currently being discussed, then, at two different levels. We have on the one hand the historian who has almost exclusively subject-oriented interests and who wants some advice on how to use computer techniques to obtain a few narrowly defined objectives without having to fully embrace the complex arguments for and against the alternative processing environments and the implications of their use in a given project. A "standard" for this clientele has to consist of specific guidance on how to accomplish a given task with the help of specific tools. On the other hand there are

[17] Such guidance exists for Κλειω in: Thomas Werner and Thomas Grotum: Sämtlich Hab und Gut … Die Analyse von Besitzstandslisten, St. Katharinen: Scripta Mercaturae 1989 (= Halbgraue Reihe zur Historischen Fachinformatik A 2); Jürgen Nemitz: Die historische Analyse städtischer Wohn- und Gewerbelagen: die Auswertung sozialtopographischer Quellen, St. Katharinen: Scripta Mercaturae 1989 (= Halbgraue Reihe zur Historischen Fachinformatik A 3); Barbara Schuh: "Von vilen und mancherley seltzamen Wunderzaichen": die Analyse von Mirakelbüchern und Wallfahrtsquellen, St. Katharinen: Scripta Mercaturae 1989 (= Halbgraue Reihe zur Historischen Fachinformatik A 4); Peter Becker: Leben, Lieben, Sterben. Die Analyse von Kirchenbüchern, St. Katharinen: Scripta Mercaturae 1989 (= Halbgraue Reihe zur Historischen Fachinformatik A 5); Steffen Wernicke and Martin Hoernes: "Umb die unzucht die ich handelt han …": Quellen zum Urfehdewesen., St. Katharinen: Scripta Mercaturae 1990 (= Halbgraue Reihe zur Historischen Fachinformatik A 9).

consultants, specialists responsible for data processing at various institutions, proproponents of various ways of teaching computing in different disciplines. For them a standard should provide the tools which would ensure that the very different solutions they collectively promote remain mutually compatible. For this group, we have to define more abstract definitions of what our kind of data processing is all about.[18] Borrowing keywords from the most recent wave of industrial PR, such definitions have to be independent of “obvious” technical solutions which get surpassed by the next wave of methodological and software evolution. A “standard” for expert users has to consist of definitions and tools which allow them to maximize their current technical knowledge and to benefit the researchers who are consulting them without precluding future technical developments.

What, then, is standardization about? Nobody needs a standard in those areas of computer processing in which only a few highly expert users are engaged. Standards are needed where many people work in slightly different ways and want to be able to communicate with one another. At least in history standards should not tell the “silly crowd” how to behave; they should define instead what all the valuable “schools” of computer usage already in existence have in common. They should, more than anything else, allow the historian well versed in one usage to drift painlessly into another without losing the data already entered and without having to relearn absolutely everything.

[18] Such definitions have been demanded for some time not just as a means of developing data processing standards but as a means of qualitatively improving data processing within the humanities. Cf. Jean Claude Gardin, “Current and Future Role of D(ata)B(ase)s in the S(ocial) S(ciences and) H(umanities) in Thomas F. Moberg (ed.), Data Bases in the Social Sciences and the Humanities 3 (Osprey: Paradigm Press, 1987), 112-6.