Manfred Thaller: Can historical information be represented outside of a graph / hypergraph / network?[1]

Rhetorical questions have usually obvious answers, so does this one: Yes, of course, people have done so for an awfully long time. The answers become usually less clear, if you specify the question a bit more precisely: Can you avoid graph-based representations when you want to go beyond what can be achieved by currently familiar technologies? Not in my opinion – though the intentionally clumsy title should indicate that we cannot simply say "endorse a graph-based data model and all problems will resolve themselves" but must define which of the properties of graph-based data models are crucial and where extensions of the current understanding of such models may be needed.

In what is called "Digital Humanities" today I recognize a tendency to mix up various unspoken epistemic as well as technical premises, which at least I find highly confusing. I feel the need, therefore, to explain first how I see the state of the art in the kind of application of information technology within historical research which I have in mind. This will be followed by a proposal for an approach to the application of information technology which in my opinion is the logical next step. Which in turn leads to a few remarks on the technical specifications for the data structures needed, which will result in the strong plea for graph-related structures implied by the title.

1. Databases in Historical Research (and some misunderstandings of Digital Humanists)

> *It says something interesting that although tools for structuring digital data have been available for historians for almost twenty years (for example: Access has been part of Microsoft's suite of programs since 1992, and FileMaker even longer than that), they have not really found a place in the repertoire of tools for most historians. Why is this so?*
>
> (Bradley 2014, 13)

John Bradley's question reads curious from a historian's point of view, at least for one who remembers that almost twenty years before this sentence was written, a monograph on how to use databases in historical research was published with a bibliography of 22 pages (Harvey and Press, 1996). In 1990, when the then MacIntosh-only FileMaker was still exotic, a systematic *Comparaison théorique des Système de Gestion de Bases de Données Relationnelles (SGBDR) Oracle, Informix et Ingres* (Pasleau, 1990) was available, directed explicitly at historians using computers. McCrank's admittedly massively overblown bibliography of ca. 5700 entries on *Historical Information Science* lists at least 500 titles relating to databases and (an overly broad conception of) the historical disciplines (McCrank 2002, 634 – 975). In the abortive conference series, with which Joseph Raben tried in the eighties to express his opinion, that Computing in the Humanities should include more disciplines than the prevalent view of ACH / ALLC at that time maintained, historical topics were among the more frequent ones (Raben and Marks 1980; Allen 1985; Moberg 1987; McCrank 1989). And of the 56 papers contained in the first volume of the History and Computing series, out of which the Association for History and Computing arose, at the very least one third was dealing with database applications (Denley and Hopkin 1987).[2]

That an extraordinarily well-informed specialist of the Digital Humanities like John Bradley manages to overlook this tradition, is an odd comment on the kind of animal the Digital Humanities have turned into under the massive onslaught of literary computing, which created the idea in recent years, that analyzing texts as stylistic objects which could be analyzed first and foremost by quantitative analyses of these stylistic traits, is the central notion of "the field". It is highly

---

[1] This text has been submitted according to CC-BY 4.0.

[2] This paragraph is also the opening paragraph of another paper of mine currently at press, entitled *Historians, Texts and Factoids* which uses the same question as starting into another direction.

meritorious of John Bradley that he has pointed out to the literary crowd, that texts can not only be the objects of study as such, but also be useful for fields of study where not the texts as such are studied, but where they are considered sources of information which must be extracted from the textual envelop. Indeed, as I have argued for some time, they various ways of analyzing historical data as sets of statistical codes, formalized databases or full texts form an analytic continuum (Thaller 1991, Thaller 2000).

Bradley – and his colleagues at King's college – use the term "factoids" for the chunks of information which according to this view are first extracted from sources for analysis (Pasin 2015). While I would not want to criticize a native speaker's English and have used the term "factoid" myself in the past, I noticed more recently that the primary definition of "factoid" in Merriam and Webster is *"an invented fact believed to be true because it appears in print"* which somehow is not the association to historical research which I consider most fortunate.

I will use the term "factlet" henceforth, therefore, when I talk about the most general information objects to be handled by information technology within historical research. That concept I consider incredibly useful to understand historical research if it is tightly bound to sources: *Factlets* lend themselves very well to the puzzle-metaphor of historical research, where a historian first and foremost tries to extract many pieces from a body of sources and then arranges and re-arranges them until they provide a convincing picture. Which may have to be redrawn many times in favor of yet another solution that integrates more pieces, dismantling previous solutions where the factlets had to be forced into the picture.

This paper makes the following assumptions[3]:

(1) Historical research consists of the interpretation of artefacts of the past in the light of today's knowledge.
(2) Such interpretation can never be objective. Nevertheless, historians have the unnegotiable obligation to strive for as much objectivity as possible.
(3) This implies the greatest possible care to separate the representation of information from the past as cleanly as possible from its interpretation.

(4) Historical sources represent tokens of data which are transmitted from the past together, individual tokens providing insights to widely different questions, often to be pursued in completely independent analyses.
(5) For many types of historical analysis relevant information is represented by tokens which cannot be interpreted unless they are considered in connection with other tokens. These may, but need not, be derived from the same source.
(6) The context of the source out of which a token has been extracted, however, must never be lost sight of.
(7) To support historical analysis optimally by information technology, we need a model to administer collections of such tokens in a way which allows the recombination of tokens without separating them from their context of transmission.

(8) Tokens taken from sources are *data*. Provisionally interpreted they become *information*. A token together with the data needed for its interpretation is called a *factlet*.

Note: Tokens *can* be texts – and will be frequently so – are *not required* to be so.

[3] Most of these are also assumptions about quite fundamental methodological positions in historical research. They are *not* discussed here but listed to present a consistent argument.

The following tries to describe an approach to support such a type of analysis. While the assumptions (1) – (8) above are probably straightforward, there are a few corollaries, not about the nature of historical work, but about the type of support an information system *should* provide for historical work, which may not be so obvious. And which admittedly is still quite Utopian – even though many current technologies look very much like enabling what I described.

It is probably quite easy to see that the sequence of assumptions above leads directly to a database system of sorts, though it may be more complex than the ones historians use currently.

Today databases are used in historical research in two different ways:

(a) Databases are in practice implicitly or explicitly considered close relatives to "editions"; John Bradley's "factoids" consider them essentially as the technical framework in which you would "edit" a source which is more strictly structured than a natural text, than a literary work or a narrative source. In this view a database is a way in which certain types of sources are prepared for the usage by the community in the digital age. And indeed, many pre-digital traditional historical projects which are dedicated to catalogues, dictionaries, thesauri, prosopographies or similar scholarly tools have in the meantime made the transition to database-oriented products. And particularly the curatorial communities' digital libraries or archives may be quite as easily referred to as "databases" just as well as "digital libraries" or "digital editions".

(b) On the other hand, in the discourse about their digital tools, historians use "my database" frequently for any kind of personal collection of snippets from source material which is administered with the help of a computer. That includes a at one extreme a Microsoft Word file with excerpts just as well as a Microsoft Access, MySQL or Neo4J database. A statement with which I try to capture the generally blurred usage within the discipline and do not want to accuse database specialists in the field of sloppy terminology.

What *both* usages have in common is, that the database is a clearly delineated tool within a much broader research process, which uses this tool only for an extremely specific purpose: frequently the administration of data from one specific source. A prosopographical – database where Bradley's factoids came from - is usually understood functionally to be a 1:1 equivalent of a printed prosopography. A Microsoft Word file with excerpts – and even quite a few personal MySQL databases containing structured and vaguely normalized records - are simply the equivalents of a stack of three-by-five-inch card stacks.

In both usages databases are:

Passive.
When you examine your database for the date of birth of a person you are working and you copy that date of birth into your most recent paper, the database does not know about it. Neither your prosopographical database, nor your private collection of excerpts will inform you that the maintainer of the database has changed that date of birth. Leave alone change the quoted date in your paper automatically, which is quite reassuring, if you base an argument upon that date and a tacitly changed one would silently invalidate your argument without you being aware of it.

Meager typed.
While it is easy to transfer a string from a database into a textual document, it may be difficult to copy a calendar date into a textual document, when it has the smallest possible semantic peculiarity – as, e.g., being a Julian rather than Gregorian one. And if a database provides a richer semantic for the database fields – the degree of certainty of a calendar date, e.g. - that is almost certainly lost.

Unidirectional.
If you are using data from a database, the database is not usually aware of it. This is behind the property of being *passive*, but is also means, that there is no obvious and simple way to feed knowledge you have gained back into the database from which you got the information which led to that knowledge.

Source Targeted.
Bradley uses the concept of factlets not so much to provide a metaphor for the *analytic* and *synthetic* stage of research (the arrangement of the parts of the puzzle), but rather for the conceptual cutting up of a source into the pieces to be arranged. According to this model almost all historical databases are dedicated to the administration of one source or a roughly similar group of sources. (If for no other reason because traditional database models favor rather rigid and regular structures.)

2. Rethinking Historical Information Systems?

Let me sketch the rough outline of a future software system that should support historical research integratedly. Though for me that leads "*back into the future"*, as I am revisiting some of the central ideas of an old concept, essentially the vision behind the historical micro studies which have been en vogue in the eighties / nineties, notably at the Max-Planck-*Institut für Geschichte* where I worked then[4], supporting those studies with a database system dedicated explicitly to historical studies (Thaller 1980, 1993, Burt 1996).

The vision ran roughly like this:

Take all registers of marriages, births, and deaths in a village / small town over two or three centuries and reconstruct all genealogical links for the population of that city. Connect *all* surviving sources of that village / small town for this time frame. Such as census lists, registers of taxation, testaments, protocols of the local ecclesiastic courts … *all* surviving sources, organized as a network of relationships between individuals mentioned in them. This should *not* be one of the collections of multimedia documents of the early years of the internet, as the *Valley of the Shadow*[5]*,* where you can browse the documents, basically using the digital medium as a cheap publication medium, but an integrated database which would be able to answer analytic questions:

Have households, where the testaments show that new, very labor intensive, root crops have already been introduced a higher incidence of illegitimate births, illegitimate children being quite convenient as cheap labor? Do protocols of the local churches show changes in the treatment of illegitimacy in the proceedings of the local church consistory as agricultural modernization progresses?

The point of these questions is of course, that to answer them we must connect information which is very widely separated. In the sources where you must look it up, as well as by the analytic tools you would use to study them: Demographic indices, proportional weights of bequeathed agricultural stores and analyses of the vocabulary used in the protocols.

The most important difference to the characteristics of databases in historical research, which we

---

4 I would like to emphasize, that these where the *concepts* pushed notably by David Sabean for some time. The published studies fell short of the ideal, not least, because at that time there simply were no algorithmic possibilities for the cheap conversion of massive collections of manuscript sources into machine readable form. Possibilities which still do not exist; seem to arise tantalizingly on the horizon, however. As examples of the results that have been achieved see (Sabean 1990), (Kriedte 1991), (Schlumbohm 1994), (Medick 1996).

5 https://valley.lib.virginia.edu/ (accessed January 12th, 2021).

have listed above, is, however, that here a database is *not* a tool to represent one *source*, but one for the dynamic construction of the evolving analysis. To come back to our metaphor: We are using the database *not* to *search for* the pieces of the puzzle, but we are using it to *arrange* them. A "family" as such does not appear in the sources, it is a preliminary step of analysis; even though the database from which the family is constructed contains at the beginning all the sources it shall create connections between. The sources remain as they are, but the analytic steps, the connections created between different sources, are highly provisional and must be open to the re-arrangement of the pieces of the puzzle at any time.

This "being open to the re-arrangement of the pieces" is considerably more challenging, than we might think at first look. Let us assume, after having worked on the kind of analysis of the stage of acceptance of new agricultural products for some time, we discover that we have misunderstood part of the mechanisms guiding local spelling variation and have therefore assigned a few individuals to the wrong families. This means obviously, that all preliminary theses about the relevance of the progress of economic modernization for behavior are potentially invalidated if the illegitimate children belong to persons connected to a family with completely different agricultural characteristics than we believed so far. If we take the argument serious, that a system supporting historical analysis should not administer the sources but support their arrangement towards synthesis, this implies that the system must have some capabilities to remember the activities triggered by the individual erroneous intermediate decisions and support backtracking from them and replacing them with corrected decisions.

Let me stop the short reminiscence of microanalytic studies and contrast the behavior of such a software system *supporting a research process*, with one *administering a collection of sources* for which we identified four central characteristics above. Short note: That I propose to look for such systems right now has a very practical reason. We are currently moving towards the systematic application of handwritten text recognition. This, together with the progress made in entity recognition, will predictably lead us in the medium future to a situation where factlets as described above – people, pieces of property, transactions – become available in very great numbers.

That traditionally databases in historical research have been targeted at the administration of sources, rather than the research processes in which their information is used, was our main argument to rethink the scope of databases in historical research. This, however, does not mean that their operations should abstract from the context of historical sources, quite on the contrary. If we follow the concept above, that individual factlets shall be taken from different sources and easily be rearranged to a representation of relationships between these factlets, we need a way to represent the individual factlets so that they fit easily together.

To extract historical factlets from historical sources and interconnect them is quite like the applications which collect short chunks of data and / or information from various contemporary information. There is one significant difference, however: Data as provided in the form of linked open data assume more or less implicitly, that they can rely upon the data to be expressed in a few basic data types which are either underlying most of modern information technology – integers, floating point numbers, all of them decimal; geographical co-ordinates connected to one of a very few co-ordinate systems – or are specified within coding systems which reflect modern habits – assuming, e.g., naming schemes which follow a first name / surname convention.

Historical sources vulnerate all of that: With non-decimal systems of measurement, unusual calendar notations, dubious projection methods in maps and many more peculiarities. If we take the principle serious which we specified above as "*This implies the greatest possible care to separate the representation of information from the past as cleanly as possible from its interpretation"* it is hard to imagine, that we propose a solution, where the person preparing a source for handling by an

information system, translates the historical data so they fit most easily into modern data types. For the described mechanism to work, we need therefore a *richly typed* environment which allows to use non-contemporary data types in a way to optimize their integration.

The most significant difference, between the way databases are currently being used and the way they would have to be used to realize the concepts described here, however, would be that they must be *active*. That is, the way in which factlets are arranged as intermediate and preliminary representations of a research process must not be understood as a sequence of database transactions, each of which is completed when it is finished. Each of those transactions should continue to be remembered in logged form, hibernating metaphorically speaking, to be revived again, when some of the steps which have led to a specific decision haves been revoked. While the idea of an active database should not create the illusion, that the system with which the historian interacts takes the *initiative* to rearrange the data, we assume that any historian who intends to change some links in the overall collection is informed about all interconnections in the data which are invalidated by this change. (Which was the one feature which was missing from the system of the eighties and nineties.)

The fourth diagnosed deficiency of the current usage of databases goes beyond our primary purpose here, the replacement of databases administering *sources* by information systems administering preliminary *interpretations* of them. We have so far assumed, that a traditional database representing a single source or a collection of such is the functional equivalent of a printed publication of that source, out of which data can be extracted to be integrated into an integrative database as discussed here. As the connection of the factlets extracted from one such source with factlets from other sources enhances the understanding of each and all of them, we expect that in future information systems such extractions do not take the form of unidirectional links between two higher order information objects, but that they are linked *bidirectionally*. Not central for our case here, but a consideration which we should keep in the back of our mind when we try to glimpse at a data model that should be able to handle the research process described here.

3. Enter the Graph.

Traditionally the term “database” raises associations of “tables”; at least it does – or did so – between ca. 1990 and 2015. If we look back at a time when “database” was not a synonym for “relational database” (and dinosaurs ruled the earth), we discover in database textbooks, that the way to decide about the most appropriate data model was then discussed, among other things, as a cost-benefit analysis leading to a decision for favoring tables or graphs as underlying data structures: the more regular the data, the more advantages for tables; the greater the irregularity, the more benefit to be drawn from graphs (Tsichritzis 1982, 28-36). For historical sources graphs always had another important advantage: For the interpretation of historical data, the context in which they appear is much more important than for contemporary data. And modelling the context a set of values provides for each other is much easier within a graph than when the values are distributed all over the place in various tables.

I always found it astonishing that for at least a decade the implicit consensus of much of the post-2000 Digital Humanities was that you would first “model” your data according to the TEI and then administer the resulting data with MySQL. While not all graphs are hierarchies, all hierarchies are graphs: “Modeling” a graph and administering it as a set of relational tables reads for me a bit like purchasing a gold-fish bowl to keep a canary bird.[6]

[6] Not to fall into the opposite extreme, I would like to point out that the NoSQL movement does not only consist of graph databases (Robinson 2013) like Neo4J (Jordan 2014) but simply acknowledges that for *some* applications relational systems are perfect; and for some they are not. (See (Redmond 2012) for a sample.) Just as a

Our case for the use of graphs could be quite straightforward, therefore. Graphs are particularly useful for irregular data; historical data are almost always highly irregular; so: use graphs for their administration. We noticed above that some of the solutions we proposed looked quite like linked open data; linked open data fit graph structures particularly well; so: use graphs for their administration. This seems so obvious to me, that I consider it rather necessary to reflect a bit on what is not so obviously appropriate about (plain and simple) graphs as supported by current software.

Are historical data truly graphs? Let us examine some of the consequences of our "puzzle metaphor" of historical research which we have introduced above.

Let *{ a, b, c }* be a set of persons in source *α* and *{ p, q }* a set of persons in source β. Our sources contain data which would make it plausible, that *each* of *{ a, b, c }* could be identical to *{ p }* as well as *{ q }* but there exists an overall restriction which allows only one element of the set *{ a, b, c }* to be also exactly one of *{ p, q }*. Less formal: According to our data each of Ann, Mary and Susan could have any of the two roles "Mother of Jim" and "Mother of Joe". Each of them could also fill both roles. But not more than one could fill each of the two roles.

How could you model such a relationship?

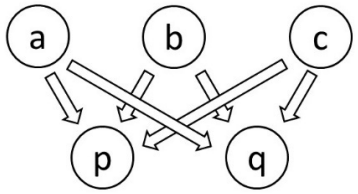

Illustration 1: A Graph

Illustration 1 is not pretty; but it maps the situation correctly. Or does it?

The purpose of our puzzle metaphor has been to illustrate the stages at which various decisions about the appropriate order in which the pieces can be arranged have to be made, and emphasize that, these decisions being preliminary, backtracking from decisions which became obsolete to an earlier stage is necessary. And there is one stage, at which we do *not* know, that there is "a connection of type <mother_of>" between *{ a }*, *{ p }* and *{ q }*. We know that that is *one* potential pair of relationships, which at the same time negates the other four alternatives in the graph. Typically, this situation arises when we start connecting the factlets: Say, we are aware that all three are candidate mothers, as their names and the dates of the two births (of *{ p }* and *{ q }*, that is) would allow. When at a later stage an additional factlet from another source suggests, that { b } is disqualified, as she was e.g., mother to another child at the same time, { a } and { c } remain candidates.

But that is at a later stage. At the beginning, all we know is, that there is *some* connection between these five nodes. Describing these changing relationships (and backtracking between them) with edges which co-exist independent of each other, though it is clear, that they exclude each other, is possible; but if, as usual at the beginning of the solution of the puzzle many alternatives are co-probable the resulting structure can be highly confusing – which is another term for error-prone.

I therefore propose to base our representation of the connections appearing through the successive solution of the puzzle not so much on a set of edges in a graph, but on a *hypergraph*[7] (Bretto 2013).

---

cautionary note that the "I need a database, so where is MySQL" should not be replaced by "I need a database, so where is Neo4J".

[7] While some texts on Neo4J report on various tricks to provide substitutes for hyperedges, these are all

A *hypergraph* can have *hyperedges*: that is, edges, which connect more than two nodes. In our example the whole relationship “three possible mothers for two possible children” would be represented by exactly *one* hyperedge, drawn usually as in illustration 2. Which in my opinion is a much better conceptual model of the situation researchers are in, when they contemplate the source situation described.

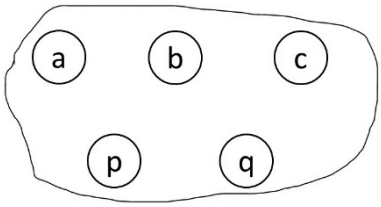

Illustration 2: Simple hypergraph

As a hypergraph where five nodes are connected by exactly *one* hyperedge – the line surrounding all five nodes – does not very much look like a graph, let us show the same set of nodes connected by *two* hyperedges. (Depicting, e.g., a stage in the puzzle, where two alternative hypotheses have developed: That either *{ b }* or *{ c }* has been the mother of *{ p }* and / or *{ q }* and independently of that there is the possibility that *{ a }* has been the mother of *{ p }*, as in illustration 3. The two hyperedges are colored to make it easier to differentiate between them. Notice that the hyperedge connecting { a } and { p } is a simple edge.

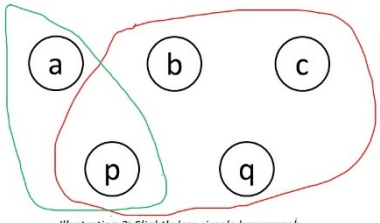

Illustration 3: Slightly less simple hypergraph

I have mentioned above that I am not aware of any production level hypergraph database; though the effort needed to implement links which behave as hyperedges, even when they are not fully supporting all formal properties of such seems to be manageable.

We are currently looking at the kind of mechanism which we would need to follow our notion of the puzzle mechanism, an environment in which the arrangement of factlets into complex hypotheses about the most meaningful interconnections can be supported. Graphs are most certainly a particularly useful way to conceptualize that. But how strictly are we bound to graph theory? In knowledge engineering it has been observed that data models derived from graphs can be useful also to analyze structures which are not strictly graphs in a mathematical sense (Giunchiglia 2004). Nested graphs, for visualization as well as just for the handling of large network-like structures are a good case in point: Here part of a property graph is contained in a node of a superordinated graph.

In the strict mathematical sense, a *graph* is only required, if we want to use analytic measures which

---

workarounds. As they seem not to be very stable – references on the internet tend to get dead – I just can point to the Neo4J online manual - https://neo4j.com/docs/ - and suggest that you search for “hyperedge”. In the version of this paper presented to the Virtual Symposium *Graph Technologies in the Humanities* 2021 I continued with: “I am not aware of any truly stable hypergraph database at production level quality now.” In a private chat I have been asked, how this related to the software offered at hypergraphdb.org. (Unfortunately, I lost access to that chat entry before I could answer it, so I cannot name the person to whom I am grateful for that hint. Apologies.) I was not aware of this system at the time. I am impressed by it, as a research endeavor and hope it draws more attention in the future, but I am not convinced, that it is at a stage where I would consider it unqualified “production level quality”.

are based upon graph theory. If we would, e.g., require a formal way to decide whether two graphs or subgraphs are isomorphic; if we want to decide upon the centrality of a node within a graph in a formal way; if we want to evaluate the formal connectivity of a graph or subgraph. As long, as none of these measures is required for analytic purposes, what we are really interested in is usually the administration of complex set of interconnections between factlets, as in property graphs. It would be interesting, how frequently the conditions for a formal evaluation of graph analytical measures are really fulfilled in historical studies. Even in networks of correspondence, certainly one of the more frequently discussed use cases, the use of a measure of connectivity or centrality would require that all edges be dimensionless, in the sense, that all of them indicate a link between two partners of correspondence with the same strength as all others, which does not seem to be very realistic. (Which is not supposed to mean, that the application of a measure of connectivity or centrality to a network of correspondence does not have a huge *heuristic* potential. Though it raises the same dangers as the application of other formal methods to data which are not really fulfilling all the prerequisites those methods require.)

We propose, therefore, to look at (hyper)graphs as a *starting point* for the data models of the "puzzle support" databases of the future, rather than as an *end point* of the necessary technical considerations before embarking upon practical experiments.

4. So what?

It is obvious, that our argument points towards Artificial Intelligence. If we intend to put pieces of a puzzle into preliminary configurations, which are successively changed as ambiguities and contradictions are resolved, it is hard not to envisage a system of rules within such a system administering puzzle pieces that would help identify such contradictions: A rule which tries to prevent a historian to assign a mother to a child who has died before its birth would be an obvious example.

For some this may be the wrong type of AI. There is a bit of confusion within the field here. On the one hand, Artificial Intelligence, as the field that tries to emulate human intelligence, has recently seen a strong move towards connectionist models, a move which has also found notice in the Digital Humanities (Buzzetti forthcoming), as part of the "Big Data" type of research. On the other hand, Artificial Intelligence, as it is currently changing industrial applications, is still very much bound to the "traditional" field of artificial intelligence, focusing on logic and rule-based approaches driving agents (like cars), automatic sub symbolic machine learning notwithstanding (Calegari 2021).

It is certainly true, that sub-symbolic (essentially: connectionist) approaches are important when we want to discover structures and relationships within data where we have no previous knowledge. In the arrangement of our puzzle pieces, the pieces and their potential relationships are very clearly defined, though. There is no abstract topic "motherhood" to be discovered within the population of a microhistorical study; the question is, who is the mother of whom. (There may very well be a topic "motherhood" to be discovered in the changes in the vocabulary used in the consistory protocols we mentioned above.)

If this is so, then we can describe the way in which some of the recent technical developments could be used for the construction of historical information systems of a new form. Presented as a set of theses on the current stage of the application of "databases" to historical sources:

(1) We are on the brink of the mass conversion of handwritten sources into machine readable form. Besides further progress in the recognition of handwritten texts, this requires a new type of software environment for the computer supported construction of databases which

connects factlets extracted from sources by entity extraction tools.

(2) To process these extracted entities or factlets, they must be represented in a form which is prepared for their integration into larger systems.
(3) While may lessons may be learned for the model for these factlets from the linked open data discussion, they must be constructed out of data types, which reflect the systems of measurement typical for historical source material.
(4) Constructing large reconstructions of interrelationships between factlets is a long and drawn-out process, during which backtracking from previous decisions must be as flexible as possible.
(5) This is easiest, if the factlets themselves are represented as small "*graphoids*" which are designed for the integration in successively more complex ones, emphasizing and supporting at all stages the management of relationships with varying probability.
(6) We speak of *graphoids*, as a data structure built from nodes and edges is most useful to create abstractions of maximum flexibility. Having said this, we want to emphasize, however, that they should be optimized for *processing*, not focus on support for any specific calculus. That will be most easily reached, if we orient ourselves from the start on the appropriate theories in their most general form, as e.g., as hypergraphs rather than graphs.
(7) For such systems relatively simple units need to be handled with maximum flexibility. Which content-defined factlets will be most useful for this is not yet clear, as they will presumably very much depend on the period and area we are working with. Much more central is a set of primitive operations, which can be used to build the emerging higher order objects out of the individual graphoids, so a definition and implementation of an experimental library of such should be a first practical step.

References

Allen, Robert F. (ed.): *Databases in the Humanities and the Social Sciences* 2, Paradigm Press 1985.

Bradley, John: "Bradley, John: "Silk Purses and Sow's Ears: Can Structured Historical Data Deal with Historical Sources?", in: *International Journal of Humanities and Arts Computing* 8.1 (2014), 13–27."

Bretto, Alain: *Hypergraph Theory*, Springer, 2013.

Burt, Janet: "Source-Oriented Data Processing. The Triumph of the Micro over the Macro?", in: *History and Computing* 8 (1996) 160-168.

Buzzetti, Dino: "Towards an Operational Approach to Computational Text Analysis", forthcoming.

Calegari, Roberta et al.: "Logic-based Technologies for Multi-Agent Systems: A Systematic Literature Review", in: *Autonomous Agents and Multi-Agent Systems* 35 (2021) article 1.

Denley, Peter and Hopkin, Deian: *History and Computing,* Manchester University Press 1987.

Giunchiglia, Fausto and Pavel Shvaiko: "Semantic Matching", in: *Knowledge Engineering Review* 18 (2004) 265-280.

Harvey, Charles and Press, Jon: *databases in Historical Research,* Macmillan 1996.

Jordan, Gregory: *Practical Neo4J*, Apress 2014.

Kriedte, Peter: *Eine Stadt am seidenen Faden. Haushalt, Hausindustrie und soziale Bewegung in Krefeld in der Mitte des 19. Jahrhunderts*, Vandenhoeck & Ruprecht 1991.

McCrank, Lawrence (ed.): *Databases in the Humanities and the Social Sciences 4*, Learned Information 1989.

McCrank, Lawrence J.: *Historical Information Science*, Information Today 2002.

Medick, Hans: *Weben und Überleben in Laichingen 1650–1900. Lokalgeschichte als allgemeine Geschichte* (= *Veröffentlichungen des Max-Planck-Instituts für Geschichte.* Bd. 126), Vandenhoeck & Ruprecht 1996.

Moberg, Thomas F. (ed.): *Databases in the Humanities and the Social Sciences* 3, Paradigm Press 1987.

Pasin, Michele and Bradley, John: "Factoid-based prosopography and computer ontologies: towards an integrated approach", in: *Digital Scholarship in the Humanities* (30.1) 2015 86-97.

Pasleau, Suzy: "Comparaison théorique des Système de Gestion de Bases de Données Relationnelles (SGBDR) Oracle, Informix et Ingres*",* in: *Revue Informatique et Statistique dans les Sciences Humaines* (1990), 183-202.

Raben, Joseph and Marks, Gregory (eds.): D*atabases in the Humanities and the Social Sciences*, North Holland 1980.

Redmond, Eric and Jim R. Wilson: *Seven Databases in Seven Weeks*, The Pragmatic Bookshelf 2012.

Robinson, Ian, Jim Webber and Emil Eifrem: *Graph Databases*, O'Reilly 2013.

Sabean, David Warren: *Property, production, and family in Neckarhausen, 1700-1870,* Cambridge University Press 1990.

Schlumbohm, Jürgen: *Lebensläufe, Familien, Höfe. Die Bauern und Heuerleute des Osnabrückischen Kirchspiels Beim in proto-industrieller Zeit, 1650–1860*, Vandenhoeck & Ruprecht 1994.

Thaller, Manfred: "Automation on Parnassus. CLIO – A Databank Oriented System for Historians", in: *Historical Social Research / Historische Sozialforschung* 15 (July 1980). doi: 10.12759/hsr.suppl.29.2017.113-137.

Thaller, Manfred: "The Need for Standards: Data Modelling and Exchange" [originally published in 1991]", in: *Historical Social Research* Suppl. 29 (2018) 203-220. doi: 10.12759/hsr.suppl.29.2017.203-220.

Thaller, Manfred: *Kleio. A Database System*, Scripta Mercaturae 1993 (= *Halbgraue Reihe zur Historischen Fachinformatik.* B 11).

Thaller, Manfred: "Historische Datenbanken. Vorteile und Probleme", in: *Geschichte und Informatik* 11 (2000), 7–25.