Semantic Networks and Historical Knowledge Management

Introducing New Methods of Computer-based Research

Author: Maximilian Kalus, M.A.

Affiliation: University of Jena

Address: Friedrich-Schiller-Universität Jena, Chair of Economic and Social History, Faculty of Economics, 07737 Jena, Germany

Telephone: +49 3641 94 33 24

E-Mail: m.kalus@wiwi.uni-jena.de

Abstract: Historical semantic networks are a new computer-based method to work with historical data. Objects (e.g. people, places, events) can be entered into a database and connected to each other in a semantic way. Both qualitative and quantitative research could profit from such an approach. Moreover, data can easily be shared among researchers. *histcross* is the name of a project in progress that implements historical semantic networks.

Keywords: historical semantic database network knowledge management *histcross* prosopographics sociographics

Introduction

Arguably, the most difficult element of historical research is the management of information. Historians have to cope with both abundant and scarce sources at the same time. Both types tend to be fragmentary and have virtually never been deliberately produced for later ages. Modern history is lucky enough to have statistical data readily available, but before 1800 things become much more difficult. In this era, qualitative research (i.e. 'traditional history') becomes much more important, while quantitative methods only play a minor role. In fact, it seems that quantitative and qualitative research do not mix well at all – many historians actually do consider themselves as followers of one faction, distrusting the other.

Because of this, it might be a bold point to state the following in a journal for quantitative history: Computer scientists have long been convinced that computers are quite apt in doing calculations and thus can solve most statistical problems. In recent years, the really hard nut to crack has been the evaluation of qualitative data. Interestingly, neither 'traditional' historians nor 'quantitative' ones have realized the potentials of these developments in computer science yet (or, if they have, the results where comparatively modest – at least from a computer scientist's point of view). Let me give some examples of qualitative computer science research that might be of interest for historians: natural language (computer linguistics), structured documents (e.g. the Semantic Web[1]), or knowledge management – and this last point is where historical knowledge comes into play.

Simplifying and structuring qualitatively complex knowledge, the interest to quantify it in some way, to make it reusable and easily accessible, all these aspects

[1] W3C. Berners-Lee 1999.

are desiterata that are not new to historians. But: Computer science is currently approaching a solution to some of these problems or at least to ease working with historical data. Consequently, this paper tries to give a brief introduction to historical knowledge management. It presents a way to both enhance qualitative research by introducing semantic networks. These also allow for qualitative knowledge to be eventually transformed into quantitative data. A work in progress called *histcross* (http://www.histcross.org) will be given as an example of what such a semantic networking database can look like.

Traditional Methods of Historical Knowledge Management

The term *knowledge management* (KM) refers to a wide range of concepts, such as corporate memories and instincts, expert systems, document managing systems, and learning organizations[2]. In a more general sense – and this is how the term will be used in this paper – it refers to methods to identify and capture knowledge in general, and to make knowledge assets available for transfer and reuse. Technical systems that implement knowledge management are called *knowledge management systems* and the knowledge assets they use are contained within a *knowledge base*. Numerous other labels have been attached to these concepts – this paper will use only the aforementioned terms in order to minimize confusion of the uninitiated reader.

On an abstract level, KM is nothing new. In fact, written matter in natural language (that is text) represents a very traditional method of knowledge management having been in use for several thousands of years. Following the definition given above, knowledge is 'captured' on paper (or parchment, papyrus or clay, for that matter)

[2] Benjamins et al. 1999, 687.

and made available for reuse. Various levels of structuring provide a more or less easy access to such knowledge. A book might contain chapters, page numbers, a list of contents and an index: All these support accessing the knowledge. It is easy to see, though, that knowledge assets contained in books are relatively costly to search. Usually, a person looking for specific information has to browse through several books and he or she still has a high chance to miss out on important pieces. Books and other written matter, while still being the predominant form to store knowledge nowadays, are not necessarily the best one.

Another traditional example is knowledge contained within (paper) files. Compared to books, the knowledge is much more structured, facilitating searches. In fact, with the advent of computers this method has inspired a large number of methods to hold structured information on IT systems. File systems and databases (as structured sets of data) are two such examples.

Historians have been using databases for quite some time, of which two flavours are of special interest: prosopographical databases and social networking models. The first type complies to KM systems in the strictest sense: They clearly define data records that require specific information (the knowledge base). Unfortunately, they are generally quite stringent and use a predefined data model. There are cases too complicated to be modelled, which clearly is not desirable at all (there are a number of not-so-nice workarounds for this). This, among other facts (i.e. the rigid data structure is simply outdated), is the reason why prosopographical base model will not be used in *histcross*, although the database is quite able to act as prosopographical knowledge base as well.

Social networks are another way to store information. Compared to prosopographical databases they possess both advantages and disadvantages. On the positive side, social networks are similar to the semantic networks described below. More precisely, semantic networks could be considered a superset of social networks. Moreover, social networks are well-established in the scientific community, therefore standard software such as UCINET[3] is easily available and has been thoroughly tested on historical data. On the negative side, they can only be employed in a narrow field. Social network theory and analysis is mostly used on quantitative data and when only a few specific variables have to be defined. The historical questions thus asked are also confined to a relatively narrow area. Additionally, fragmentary data sets and generic knowledge are not exactly their strengths. Still, they form a good base to start defining data models from.

Summarizing, both prosopographical and social data sets possess interesting properties, but are too narrow to comply to the need of working with data on a broader base. Wouldn't it be interesting to merge the advantages of both concepts? In fact, several relatively recent developments have made this possible. In the following section another approach will be introduced.

Theoretical Background: Frames and Semantic Networks

In the wake of artificial intelligence research, scientists soon were required to cope with the question of how a mind retains information and how it would be possible to represent such information in an artificial system. Several structural proposals were made, but one received special attention and still plays a role today: In 1974,

[3] Analytic Technologies.

Marvin Minsky[4] proposed a new theory based on memory structures called *frames*. In essence, frames represent pieces of information that contain attributes called *slots* or *fillers*. An example: A frame could represent a specific human. In this case slots could be first and last names, birth date, hair color, etc. Frames do not necessarily model specific objects but can also stand for prototypical and deducted knowledge. A generic frame *city* (as depicted in figure 1) could contain prototypical slots like *walled = yes*, *bishop's seat = no* and *number of households = unknown*. A *bishop's city* frame could inherit that information, but change the prototypical *bishop's seat* to *yes*. Finally, a specific city might have *walled = no* and *number of households = around 5,000*. Frames are machine-usable formalizations of concepts or schemata. Frame theory has close ties to object oriented programming, which could be regarded as a specific implementation of frame theory (historically, they developed side by side).

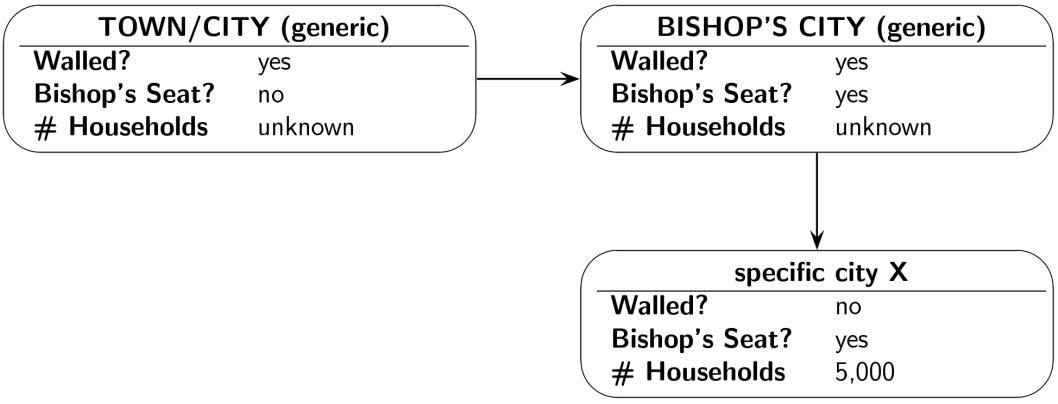

Figure 1: An example of frame based knowledge.

Frames are a useful way to represent knowledge. In addition, it is quite easy for humans to understand since it does indeed model some of our own memory structures. The *histcross* database has partly been inspired by frame theory: It can handle different types and there are certain slots that expect data.

[4] Minsky 1975.

Alas, Minsky's concept is monolithic – every information related to a certain object is stored within a frame. Although this is operationally easy to implement it is somewhat limiting. In recent years, a more flexible approach has grown in prominence: semantic networks. In a formal way, a semantic network is a directed graph consisting of vertices which represent concepts and edges which represent semantic relations between the concepts. Since it is a graph, all the concepts and lemmas of graph theory[5] can be applied. Thus, its depth, shortest path, or distances between nodes, the degree of a node, etc. – all these can be used to explore semantic networks. Structurally, semantic networks differ from frame theory in how information is stored: Frames store slots and other information within the frame, while in semantic networks these are represented by semantic relations between nodes (See figure 2 as an example of a semantic network).

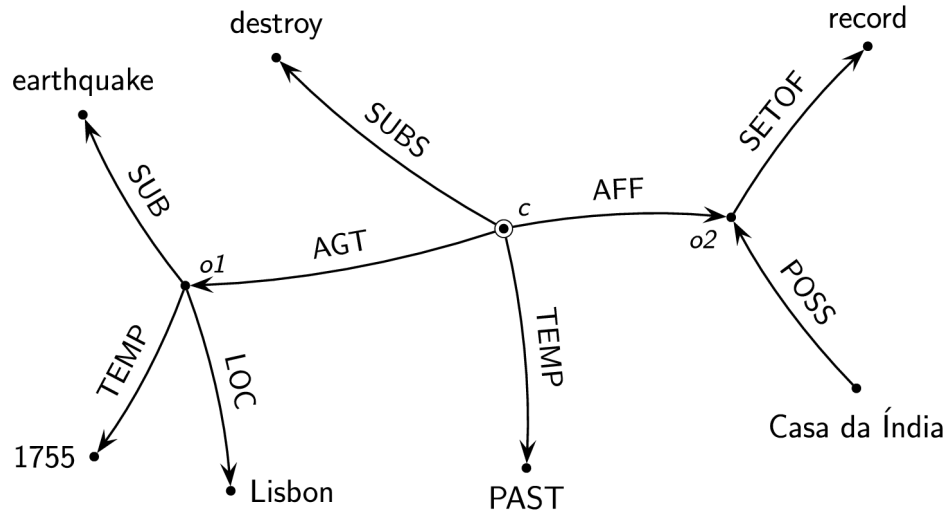

Figure 2: A semantic network in MultiNet[6] syntax.

Based on linguistic theories, semantic networks try to meet both mind structure and natural language modelling requirements. They are generic concepts to describe knowledge which can be analyzed both qualitatively, by browsing the net and

[5] More on graph theory: Bondy/Murty 1976.Diestel 2005.
[6] Helbig 2006.

interpreting nodes and edges, and quantitatively, by giving nodes weight numbers, by counting relations, or by applying statistical algorithms or automatized inferences. In fact, social networks – they have been used by historians quite extensively – are just a highly specialized form of semantic networks. Consequently, semantic networks can describe formal and informal information in a very generic (and still operationally useful) way.

Both concepts can be merged: If frames are regarded as nodes in a networks, both frame and semantic network theory can be merged. *histcross* is an implementation of this.

Before describing *histcross*, it should be noted that there are other possibilities to manage knowledge. One such concept has already been introduced in this journal in 2004, videlicet XML[7]. XML is an extremely versatile concept that has become quite popular since its official recommendation 1998. In short, XML is a language to generically describe any type of data in an easy and both human and machine readable way. It can be used to model formal data, natural language and meta-information. In connection with KM it has attracted attention in combination with structured document theory which provides a means to separate content (data) from layout. Without going into further details, XML has tremendous possibilities but one grievous drawback: XML is slow. Because of this, XML is generally not used to store live data, but (only) in data exchange and description (pre-rendering) scenarios, both of which are not performance-critical. Thus, the role of XML in KM is relatively small and *histcross* will not use XML either (except, possibly, for the aforementioned data exchange, of course).

[7] Schaefer 2004. Spaeth 2004.

Introducing Historical Semantic Networks: histcross

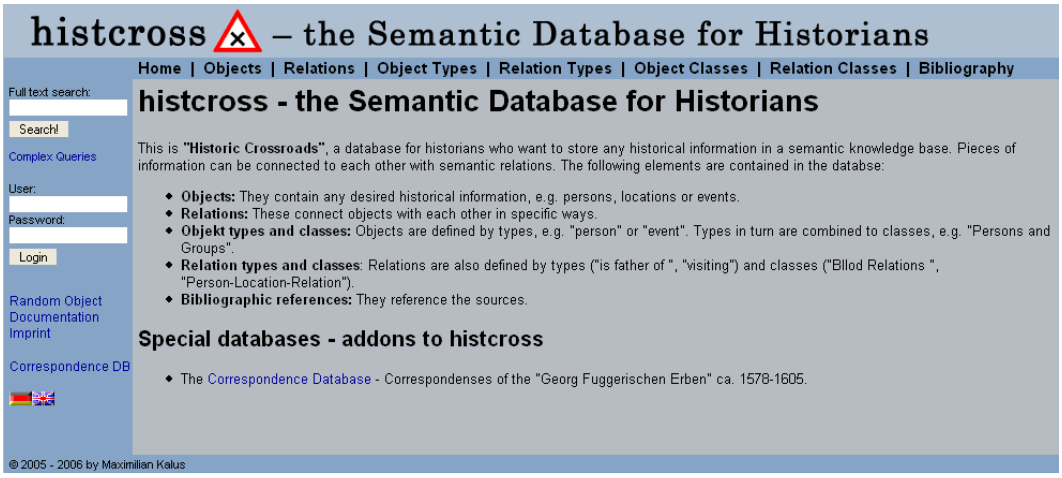

Figure 3: Screenshot of *histcross*.

Historic Crossroads or *histcross* is the name of a database that combines a frame based model with a semantic networking one in order to create and maintain historical knowledge (see figure 3). *histcross* follows three principles:

- **Simplicity:** Complexity slows down the person entering or changing data. To speed up working with the database and to keep the learning curve low, the database has to be comparatively simple.
- **Generic data:** The data model of *histcross* has to be flexible. Users can define their own types and create their own inference rules.
- **Accessibility:** *histcross* is web-based. Individual implementations can be accessed world-wide without the need of installing extra software (except for the browser, of course).

Complying to graph theory, there are two principal data structures: nodes and relations. Nodes – called objects in *histcross* – represent historical events, places, people, goods, groups, concepts and the like. Each object has a number of fields that may contain data:

- **Type and class:** Types and classes help to classify objects. Each object is member of one type. Examples for types are *person*, *city*, *village*, *event*, *ship* or *trading good*. Each type is member of a class: Cities and villages are both types of the class *location*.
- **Title:** This is a label given to the object, a person's name for example. Alternative spellings would generally go into the comment section.
- **Comment:** A text of any length that describes the object. Usually, natural language information, alternative spellings of the title, parts of excerpts are kept here. In short, comment is a catch-all field for information.
- **Start and stop date:** Naturally, dates are a central theme in history. A number of ontologies[8] exist to logically describe time and time intervals. *histcross* uses the common interval model: There is a start date b(t) and a stop date e(t). A point in time is described as a time interval where b(t) = e(t). Both entries are optional, of course. The user can just enter a start date or a stop date or none at all. Moreover, there are several granularity options: First, the user can just enter the date as year, as year–month, or as year–month–day (which is the smallest amount of time *histcross* can handle). Secondly, a granularity option allows the settings *exact*, *circa*, or *unknown/unsure*. Although the system does not (yet) interpret this additional qualifier, it forms an easy way for the user to see the reliability of the information. The last option is the calendar setting. At the moment, *histcross* can handle the Julian and Gregorian calendars (the later one being called 'automatic', because it automatically switches the calendar system on October 14th, 1582). This setting is important when specific dates are

[8] van Benthem 1991.

searched or compared. Internally, *histcross* keeps dates in the Julian Day Number format (this is the number of days having elapsed since Monday, January 1st, 4713 BC in the proleptic Julian calendar).

- **Icon:** Each object may have one icon attached from the icon database which shows up near the label. This makes it easier to visually recognize an object.
- **Bibliographic entries:** A list of bibliographic entries (from the bibliographic database of *histcross*) can be attached to each item.

Obviously, the data model is relatively simple. Yet, it is possible to gradually add data to the database during research. An example object filled with data can be seen in figure 4.

Label	Madre de Deus de Guadalupe
Type	Ship
Class	Objects
Start date	1596 [ca]
End date	1596 [exact]
Description	ship of the pepper trade; capsizes in Cochin; ...
Bibl. Ref.	Boyajian: Trade under the Habsburgs, p. 27.

Figure 4: *histcross* object example.

The second structural data element of *histcross* is the connection between objects. Relations, as they are called in the database, can each connect two objects in a semantic way. The basic data is similar to that of objects with the following exceptions:

- **No titles:** Relations do not possess titles. Rather they are defined by their type only. This could be something like *is mother of*, *is located in*, etc.
- **From-object and to-object:** These specify the two objects that are connected to the relation. It should be noted that the relation is directed to form relations like *A–is mother of–B*. All connections are bi-directional.

All the other fields are exactly the same. A relation can contain a start and a stop date, a comment and bibliographic entries. Because of this, the actual difference in handling objects or relations is not that big.

A typical (but simplified) semantic historical network described by *histcross* can be seen in figure 5. The shape of the nodes depicts the type and the class as shown in the legend. The example of Octavian Secundus Fugger shows the use of different date granularities: His birth date is unknown, while the date of his death is. Likewise, it is not known when exactly the *Fuggerische Erben company* started their operation in Goa, but estimates point to around 1587. This is depicted by the question mark after the start date of the relation *Fuggerische Erben company–operating branch in–Goa*.

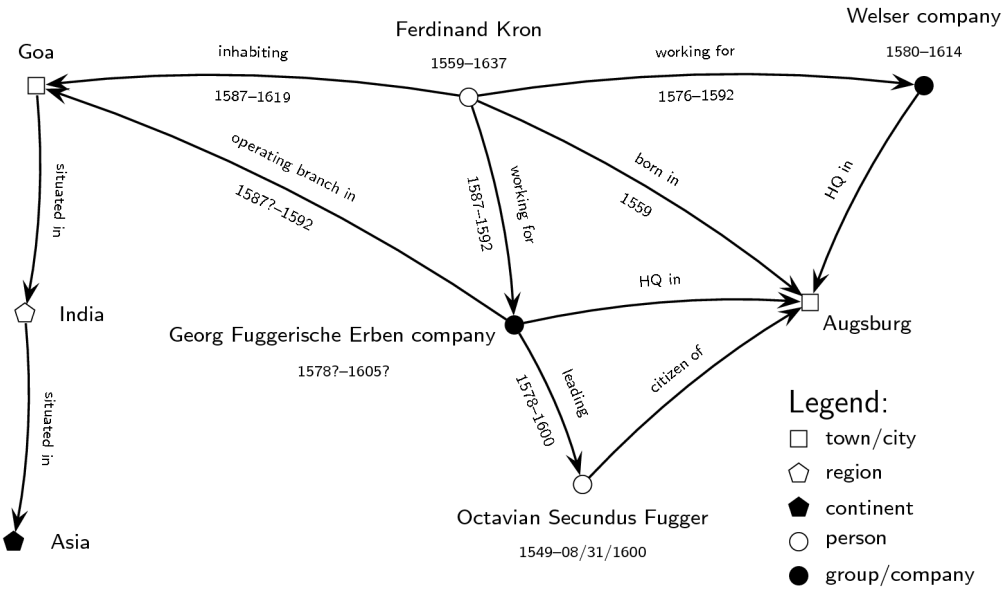

Figure 5: Historical semantic network example.

Before some implementation details of *histcross* will be introduced, it should be stated that the database still is a project in development and that there are many more possible additions not yet implemented. To name some: external references (links to other databases in the web), automated plotting of networks, sociological

or geographical analysis, or the quantification of relations (adding numbers and measurement units to relations).

Operations on the Knowledge Base

After this short introduction of *histcross*, what is the advantage of using a historical semantic database? The main benefit surely is the possibility to accumulate knowledge in a structured way during research. For this purpose, semantic networks are not the only method, but *histcross* offers a standardized tool which can be used by individuals or by a group of researchers. Moreover, it can be searched fast (there is a full text search). The data compiled in the knowledge base can be made public (on the Internet) or at least available for certain people (in an intranet). Last but not least, it is closer to the sources than a scientific paper by possibly offering the raw data. This approach saves time and research resources and is a further step to networking within the scientific community itself.

In short, *histcross* enables the user to easily create and manage historical knowledge bases step-by-step. To double back to the introduction: The user can actually enter qualitative data in a standardized form. He can thus make this data quantifiable. Additionally, besides the full-text search, the database boasts a query editor to create complex searches in the database. For example, *histcross* can answer questions like: 'Show me all the Italian merchants that traded with pepper in India after 1549.' Naturally, in order to get meaningful answers to such a question, the data has to contain objects like *India*, *pepper*, *Italian* and the merchants' objects that are related to all those objects. However, this small example shows the

opportunity of semantic systems: The user is not confined to full-text searches, but has the possibility to undertake complex semantic queries.

This semantic quality of *histcross* has many implications. Quantification of data has already been mentioned. Connected to this are certain types of operations: Counting and measuring the 'geography' of the network, analyzing 'central' and 'peripheral' nodes in the network by extracting certain subsets from the knowledge base. Consequently, sociographical analyses become possible. One of the more powerful features will be elaborated below: the possibility to add automatized rules to create new information.

In order to simulate artificial intelligence it is necessary that a system can somehow deduce new information from existing knowledge. This is generally done by implementing so-called *inference rules*. *histcross* uses a relatively simple inference engine, which can nevertheless handle the largest part of the inference requirements that might be wanted in historical research. Specifically, the application uses several forms of deductive chains that approximate predicate logic. For example, if we know that Goa is in India and India is in Asia, we can assume that Goa also is in Asia. Logically, we could write is_in(Goa, India) $\wedge$ is_in(India, Asia) $\Rightarrow$ is_in(Goa, Asia), or – as a general rule: P1(x, y) $\wedge$ P2(y, z) $\Rightarrow$ P3(x, z). *histcross* can also implement variants of this formula, e.g P1(x, y) $\wedge$ P2(z, y) $\Rightarrow$ P3(x, z). P1, P2 and P3 are not necessarily the same predicates, but will be semantically close in most cases. Some examples:

is_mother_of(x,y) $\wedge$ is_father_of(z,y) $\Rightarrow$ is_husband_of(z,x)

is_father_of(x,y) $\wedge$ is_father_of(x,z) $\Rightarrow$ is_sibling_of(y,z)

is_citizen_of(x,y) $\wedge$ has_confession(y,z) $\Rightarrow$ has_confession(x,z)

It has to be stressed that these inference rules have to be optional implications rather than mandatory ones. The father of a child is not necessarily the husband of his or her mother (in case of an illegitimate child), a citizen of a city can follow another confession than the city's official one. Because of this, whenever a user creates or changes relations, a list of *possible* inferences are presented to the user to choose from. As such, inference rules speed up data acquisition and support finding new inferences (both come very handy when entering genealogical data, for example).

Conclusion

Up until now, semantic databases have received little attention in historical research. The reasons for this are two-fold: Attributable to their somewhat different perception of time, historians trust traditional data storage matter (paper) more than modern ones (the problems of archiving electronic data are obvious). Moreover, most historians (mostly as computer amateurs) only possess a vague idea of the possibilities of computing. On the other hand, computer scientists have little concept of the qualitative problems historians have to cope with – unlike linguistics, history lacks a well-established field of computer history, in which these concepts could be advanced systematically. Also, semantic concepts in computing are still relatively new and are chiefly used within a comparatively close-knit community of knowledge management specialists.

By this short introduction of *histcross* as a historical semantic database, I intended to breach this barrier and give a brief glimpse into the possibilities of these

relatively recent developments in information technology. Knowledge management combined with historical research is a matter worthwhile to be explored further.

Of course, much more could be said on the matter, on theories and implementation aspects, but this short paper has shown the key elements: Prosopographics can form a base for historical semantic databases, but the later go a step further and allow for new dimensions in research. Databases like *histcross* can be used in historical sociology (like the analysis of elites or clienteles), historical geography, reconstruction of source material, analysis of communication networks or that of other networks (like my own research project that attempts to reconstruct merchant networks in the European–Indian spice trade of the 16th century). Naturally, a semantic database is not the ultimate tool to work with historical data, but it can support research work tremendously by automatizing certain research aspects and supporting easy search and addition of data. To conclude, semantic networking achieves something that does sound odd at first glance: It does offer a possibility to merge qualitative and quantitative aspects.

Bibliography

- Analytic Technologies: UCINET 6 Social Network Analysis Software. http://www.analytictech.com/ucinet/ucinet.htm.
- Benjamins, V. R., et al. 1999. $(KA)^2$: Building ontologies for the Internet: A mid term report. *International Journal of Human-Computer Studies* 51: 687–712.
- van Benthem, J. 1991: *The Logic of Time: A Model-Theoretic Investigation into the Varieties of Temporal Ontology and Temporal Discourse*. Dordrecht: Kluwer.

- Berners-Lee, T. 1999: *Weaving the Web. The Past, Present and Future of the World Wide Web by its Inventor*. London: Orion Business.
- Bondy, J. A., and U. S. R. Murty. 1976. *Graph Theory with Applications*, London: Macmillan. http://www.ecp6.jussieu.fr/pageperso/bondy/books/gtwa/gtwa.html.
- Diestel, R. 2005: *Graph Theory*. Heidelberg: Springer. Graduate Texts in Mathematics, 173. http://www.math.uni-hamburg.de/home/diestel/books/graph.theory/.
- Helbig, H. 2006: *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.
- Minsky, M. A. 1975: A Framework for Representing Knowledge. In *The Psychology of Computer Vision*, edited by P. Winston, 211–277. New York: McGraw-Hill (See also: MIT-AI Laboratory Memo 306, June, 1974, http://web.media.mit.edu/~minsky/papers/Frames/frames.html).
- Schaefer, M. 2004: Design and Implementation of a Proposed Standard for Digital Storage and Internet-based Retrieval of Data from the Tithe Survey of England and Wales. *Historical Methods* 37/2:61–72.
- Spaeth, D. A. 2004: Representing Text as Data: The Analysis of Historical Sources in XML. *Historical Methods* 37/2:73–86.
- W3C: *Semantic Net Activity*. http://www.w3.org/2001/sw/.