Semantic and Thematic Navigation in Electronic Encyclopedias

Henning Lobin and Andreas Witt

Bielefeld University, Germany
lobin@lili.uni-bielefeld.de

1 Introduction

In the field of electronic publishing, encyclopedias represent a unique sort of text for investigating advanced methods of navigation. The user of an electronic encyclopedia normally expects special methods for accessing the entries in an encyclopedia database. Navigation through printed encyclopedias in the traditional sense focuses on the alphabetic order of the entries. In electronic encyclopedias, however, thematic structuring of lemmas and, of course, extensive (hyper-)linking mechanisms have been added. This paper will focus on showing developments, which go beyond these navigational structures. We will concentrate on the semantic space formed by lemmas to build a network of semantic distances and thematic trails through the encyclopedia.

2 SGML data

The basis on which the information in an electronic encyclopedia is represented should be an SGML or (if possible) an XML data structure.[1] Usually, publishers of electronic encyclopedias already have a basis of existing texts and pictures. If this data is not already present in SGML form, the data must first be converted into this general annotation format.

SGML allows for structuring a textual as well as a non-textual database in different ways. It is widely agreed, that SGML should be used for annotating information that is content-specific, rather than information specific to printing or formatting. The question is, however, what exactly does content-specific information mean? The SGML Document Type Definitions primarily used for conventional publishing (e.g. ISO12083), but also those designed for electronic publishing (e.g. HTML) focus mainly on annotating information specific to document structure, i.e. headings, paragraphs, etc. On the one hand this kind of information is independent of the type of media of the resulting text. On the other hand, there still exists a strong print metaphor. A real media independent representation format should also take semantic information into account (e.g. Detroit is a big city located in the USA). Annotating detailed semantic meta-information is the basis for several applications:

- for using semantic linking it is necessary to annotate semantic information;
- even more information is necessary in respect to an automatic
- text-design based on heuristics and
- generation of excerpts
- as a basis for generating visualizations dynamically.

The following sections describe all of these points. Without going into the details of the organization of an SGML Document Type Definition just one more important aspect should be mentioned. Since an encyclopedia consists of different types of lemmas (geographic lemmas, information about persons) it is necessary that a DTD covers a large semantic structure.[2]

3 Semantic Annotation

Encyclopedias are an invaluable source for automatic content-driven evaluation. The encyclopedias themselves benefit from this evaluation. One useful exploitation of the semantic

annotation is the computation of semantic distances between lemmas. The semantic distance measures the degree of the semantic nearness of two lemmas. For instance, the semantic distance of “Rome” and the “Vatican City” is quite small and the semantic distance of the lemma “Rome” and the lemma “Elephant” is relatively large. In general the semantic distances are computed statistically, although it is also possible to provide the semantic distance manually by building up a semantic network, for instance (cf. Sjoka 1998). Encyclopedias are an ideal source for a statistic measurement of the semantic distance. Reasons that make them suitable for this computation are:

- their coherent structure
- nearly equally weighted lemmas
- largeness.

With this data it is impossible to accidentally compute an incorrect semantic distance. For example, if a textual base is used, which is too small, it might be possible to compute a semantic nearness of “elephant” and “Rome” because of the fact, that there is a monument of Bernini in Rome portraying an elephant.

There are several strategies for computing the semantic distances of the words in a given text corpus. Most of these strategies make use of the distances between the words in the text. The main drawback of this approach is that the resulting semantic distances depend on the topic and even on the style of the text corpus. In the case of encyclopedias, the preconditions for computing semantic distances are far better. In an encyclopedia provided by a professional publishing house, the lemma of an entry clearly indicates the semantic center of processing. Furthermore, each entry is balanced with respect to important and unimportant information concerning the lemma. Therefore, the analysis of articles in encyclopedias can be achieved using less complex methods and algorithms and at the same time yield better results than the semantic analysis of an arbitrarily chosen text.

4 Exploitation of Semantic Annotation

4.1 Explicit linking

An important aspect of semantic distances is the possibility of a (possibly 3-D) visualization. Using this technique involves building up an association space of the lemmas. The user of this visualization is able to “fly” through this space and can “land” at a lemma, whose associated text will then be shown. By means of this technique the semantic distances are transferred to spatial relations, which are much more intuitive for the user than the classical chains of pointers. Additionally, in a visualized semantic network the user has also the advantage of being able to foresee semantic relations. The computer program *Perspecta Server/Viewer* from the company Perspecta allows for a specification of such spaces on the basis of SGML documents. (cf. Holtzman 1997)

An option for newer digital encyclopedias should be the ability to access the lemmas using these software-tools.[3] Accessing of this type allows the user to explore and traverse the net of knowledge. Therefore, a goal for newer digital encyclopedias should be to completely link all lemmas with one another. The problem is, however, being able to administrate a semantic network with the size of an encyclopedia which has hundreds of thousands of lemmas and, consequently, millions of interconnections between them. These problems are primarily not technical ones, but rather ones concerning practicality. Afterall, constructing such a semantic network requires a lot of man power. Given the situation that an already existing textual base should be converted to an interconnected structure of lemmas, an incremental approach is preferred. That means that, as a first step, the lemmas should be linked with each other in a loose way and, later on – little by little – all connections of secondary importance must be added.

4.2 Automatic Text-Design based on Heuristics

To be able to use the electronic articles in a flexible way it is desirable to implement an automatic design of the text. This allows the user to specify the form of the information individually with parameters, for example the size of the text or the location of pictures. As a consequence, the provider of the information cannot be sure of how the information will appear on the user's computer screen. A semantic structuring of the texts and their including non-textual information enables a rule-based solution of this problem. Furthermore if the user wants to get a printed version of these texts, the preparation of an appropriate text can benefit from an evaluation of the semantic context.

4.3 Generation of Excerpts

The improvement on the flexibility of the information's presentation does not stop at the borders of an article. If the user wants to explore a broader area of knowledge, he or she can be supported by the provision of larger excerpts from the lexicon. Given the example that somebody would like to explore the theme *Baroque Music* systematically, it is unavoidable to leave the article of the lemma and to enter different articles (e.g. via *Baroque Music* through *Bach* to *Mathematics*). This well-known problem ("getting lost in cyberspace") is manageable if the user gets an excerpt of the lemma. This excerpt might include all connections to the article. Those articles deviating from the theme are marked in a special way. So, on the one hand the user still has the option to explore other topics, but abandoning the first topic will be the result of his or herdecision.

Excerpts consisting of many small pieces of information that make up a hypertext are called closed hypertexts. The advantages of closed hypertexts compared to open hypertexts are that the user can more easily maintain an overview of the topic, and the task of obtaining a coherent model about the knowledge can be solved faster. If the user chooses the mode of excerpt, rather than a "classical" presentation-mode of the articles, the newly generated text should be a hierarchical text, containing headings, sections etc. Of course the user should also be able to print out the text of the automatically generated article. Especially in this mode of presentation there might be quite a large difference between the printed version which contains more text, and the version on the screen containing more hyperlinks.

4.4 Implicit Linking

One of the characteristics of an encyclopedia is that de facto every term is included in it. Therefore the user may expect that every noun in the text is "clickable", i.e. is connected to further information. Linking of this kind should not be made explicit in the way described above for two reasons: First, this would lead to unwanted redundancy and, secondly, an update of the database containing the encyclopedia might lead to incorrect explicit links. Therefore, a (semi-)automatic approach should be taken.

A table is constructed including all nouns in all different forms. To each noun exactly one lemma is specified at which should be pointed to by a link. This approach offers two advantages over alternative methods. At first, a table offers the possibility of a fast access to the data, especially in comparison to a computation on demand, and, secondly the preparation of a fixed table containing the connections described allows the user to enter a connected keyword without having to choose it from a list of several keywords beforehand.[4]

The extraction of implicit connections could even go beyond the connection of just keywords. Within an electronic encyclopedia a complete linking of all words is possible. That means that not only the nouns will be connected to keywords, but also all other words included in the articles. A lot of these links simply point to dictionary entries, but quite often a linking to keywords will be possible, too. For instance, it might be desirable to link the adjective *Dutch* to the keyword *Netherlands* or the word *playing* within some semantic contexts, e.g. in an article about *Arthur Rubinstein*, with the keyword *music*. Despite all of this, it should be emphasized

that implicit linking of words should not and cannot substitute explicit linking described above.

5 Presentation

Let us first look at some sample types of articles:

- *Countries*, *Capitals*
  The articles that highly depend on visual information (photographs, maps) such as ones that concern countries or capitals cannot be represented with hypermedia in this way. Instead, the photos should be arranged as tours of the cities, and the maps expanded to ‚clickable maps' with which the user can retrieve pictures or further information about the city or region.
- *Picture-Maps*
  These elements of information must also be realized in a special way appropriate for hypermedia. So-called trails, which provide the user with information step by step are suitable for this. It is important in both cases that the information contained in explanations of diagrams and legends is connected to the article so that it can also be accessed by search procedures.
- *Visual Information*
  The problem with visual information is that important parts of the article are transferred to the grafical representation. There are two possibilities to make the information accessable (for example for a search engine): The information contained in the grafical representation is stored independently from the grafics themselves for example in a separate database, from which the visual information can be constructed when putting a product together. The second possibility would be to maintain visual information and non-visual information separately which makes making change more difficult but at the same time allows freedom in the design of the grafic.
  It can generally be said that new elements of presentation that will be used also for the printable version of encyclopedias offer nowadays at the same time many more possibilities for the development of hypermedial applications since the structure of the articles which is mainly oriented towards visualization and text design compliments the hypermedial medium.

5.1 Information splitting

The presentation of information in the form of hypertext (units of information are connected with one another by hyperlinks and can be traversed in any number of ways) can not be seen as a completely different type of text with respect to traditional linear text. Even the nodes in a hypertext are normal linear texts and even linear texts can contain non-linear expansions, for example links, directories, etc. Both forms of presentation have advantages and disadvantages for the user. On the one hand, hypertexts make associative navigation in the realm of information possible, but on the other hand getting a coherent overview of a certain subject is made more difficult since the user is often unsure of the boundries and scope of the area. With linear text, however, the author can explicitly determine the order of the presentation of the information, which makes creating a coherent overview of a subject somewhat easier. However, the price to be paid is less flexibility in dealing with the information.

Hypermedial systems of the future will no longer deal exclusively with just one of these techniques of presentation. Both the size of the units of information connected with hyperlinks and the characteristics of presentation must be able to adapt to the user's preferences.
That means that for the presentation of the articles in the encyclopedia a version with small units of informtion and many links as well as a version with larger units and fewer links must be available. The user must be able to choose between these two versions. A somewhat larger article, like the one about Johann Sebastian Bach, should be arranged in smaller parts whose links can be accessed by a navigator according to the first type of presentation. A vast number

of links refer to other links, especially the relevant keywords, for example, Bach's sons, Leipzig, other Baroque composers, etc. The rather traditional presentation would leave the article in approximately the same form as it appears in the print version of an encyclopedia and would additionally contain a hypertextual table of contents. The most important links could appear in a special section quite similar to a bibliography. Both forms of presentation can be realized if the articles exist in a structured form. The functional semantic indication of individual sections of text can be interpreted as a link as well as the boundry of an ordinary segment of text.

5.2 Flexible text presentation

Another alternative (an intermediary form that is possible due to the structural marking of the articles) is to use definitions that appear between brackets or glosses. Instead of just indicating a link typographically, for example:

> ... Johann Sebastian Bach's four sons were important musicians and composers, especially Carl Philipp Emanuel and Johann Christian. In 1845 the last male descendant of Bach died ...

the part of the article containing the information that is referred to can be blended in as a short piece of information:

> ... Johann Sebastian Bach's four sons were important musicians and composers, especially Carl Philipp Emanuel (called „der Berliner" or „Hamburger Bach", 1714 to 1788 ➢) and Johann Christian (called „der Mailänder" or „Londoner Bach", 1735 to 1782 ➢). In 1845 the last male descendant of Bach died ...

The arrow indicates the starting point of the links that lead to the complete articles about the two sons. The glosses show another solution where the definitions taken from the article that was referenced appear as additional text in their own column next to the original text which remains unchanged:

> ... Johann Sebastian Bach's four sons
> were important musicians and composers,
> especially Carl Philipp Emanuel and Johann ← called „der Berliner" or „Hamburger Bach", 1714 to 1788
> Christian. In 1845 the last male descendant ← called „der Mailänder" or „Londoner Bach", 1735 to 1782
> of Bach died...

The text in each commentary can then serve as the starting point for links to the corresponding articles.

5.3 Visualization of semantic networks: The Brain

The semantic structure is a network that shows a lemma surrounded by other lemmas in the near semantic environment connected by arcs. As an interface we have adopted an innovative approach to desktop management, called “The Brain” (by Natrificial; see Fig. 1).

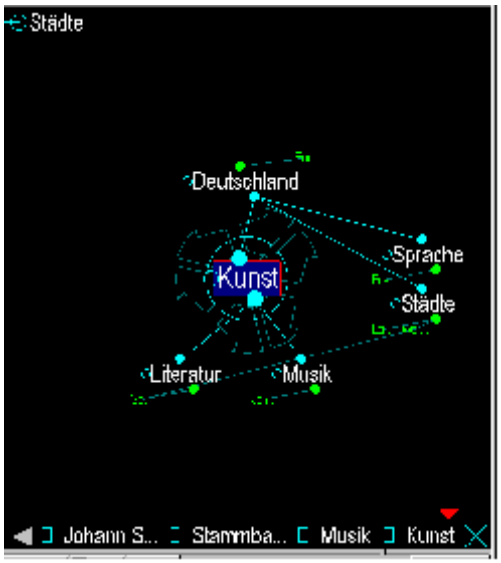

Fig. 1

This interface software allows the user to travel through the semantic space simply by clicking on the lemmas surrounding the lemma being focused on. If the user activates the focused lemma, the corresponding article is displayed in a separate window. Traveling through the encyclopedia semantically enables the user to investigate complete knowledge areas much better than hyperlinking mechanisms do. The user can “see” the knowledge area as a network structure and can, therefore, avoid getting lost in hyperspace. (cf. 4.3)

5.4 Hotspots and Hypertrails

Regarding the use of pictures, for example the map of a city, or larger graphical overviews, the user already expects to find links on certain areas of the picture that he can use to jump to explanatory texts. Within the framework of an SGML annotation for the substance of an encyclopedia this can be achieved by the use of so-called *Hotspot Architectural Forms,* which allow regions to be defined for numerous picture formats independently of their scaling. These regions can be used as bidirectional links or in other words, to the region and from the region. (cf. Rubinsky and Malony 1997)

Another still relatively new mechanism we are using are hypertrails. Hypertrails allow for a traversing of the knowledge base guided by certain criteria. Hypertrails are predefined paths through the network of information and can for this reason be constructed according to didactical perspectives. Hypertrails are implemented in the form of meta-links that connects the relevant information without being seen when normally used. In addition to that, hypertrails may also contain special connecting or explanatory comments that make the journey down a hypertrail similar to a tutorial (see Fig. 2).

An especially interesting aspect of hypertrails is that they can practically guide the user through the internet. The current information available online can be connected to the ”more stable” information in the encyclopedia in this way, for example a calendar of events, or political information. Of course the hypertrail through the internet can also pass through the publisher’s

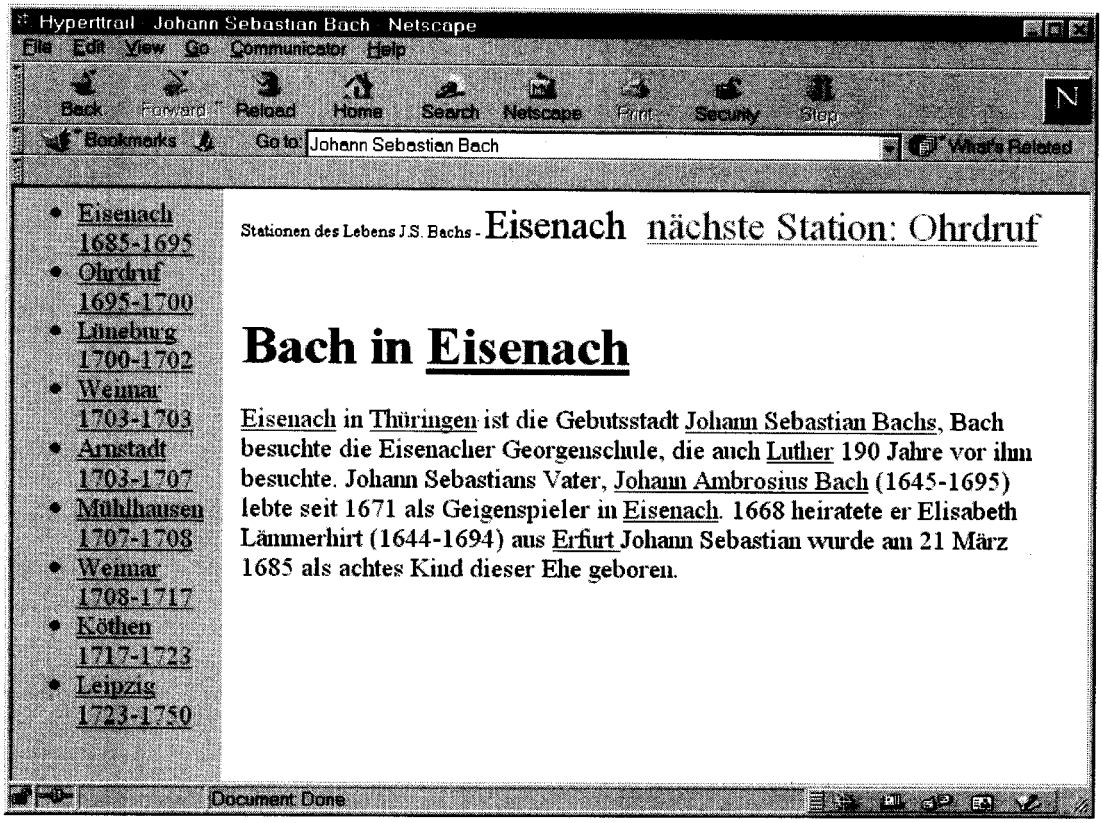

Fig. 2

own server where new information and current products concerning the topic can be found.

In hypermedial systems it is normally possible to return to an anchor of a previously visited link by clicking on the „Back" button. However, it is not possible to see a list of all the units of information that have a link pointing to the current unit of information. A relatively simple evaluation of the substance, however, allows us to obtain this functionality. For example, a categorized list of all the articles containing a link to the article about Bach could be added to the article itself.

6 Outlook: Putting together your „own lexikon"

The most demanding adaptation the product may be expected to make would be to assist the user in constructing his or her own sub-lexikon. A user might take advantage of this functionality when preparing a class presentation or talk. The user can choose certain articles from the lexikon and adapt the methods of navigation available in the entire lexikon to the needs of the sub-lexikon. Finally, the keystone of such a module is the necessary support for editing the articles which could be something similar to an interface for a word processing or presentation program (for example *Microsoft Powerpoint*).

Bibliography

Sperberg-McQueen, C. M. and Lou Burnard, *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Chicago and Oxford: Text Encoding Initiative.

Holtzman, Steven, *Digital Mosaics*. New York: Simon & Schuster, 1997.

Rubinsky, Yuri and Murray Maloney, *SGML on the Web*. Upper Saddle River (NJ): Prentice Hall, 1997.

Sojka, Petr, Publishing Encyclopedia with Acrobat using TeX. In *Electronic Publishing 98, Conference Proceedings*, Budapest, 1998, pp. 217-222.

ISO12083 International Organization for Standardization. *ISO 12083:1993(E) Information and documentation - Electronic manuscript preparation and markup.* Geneva: International Organization for Standardization, 1994.

Notes

1. The choice of XML might allow for the use of special means of the XML Linking Language(s), i.g. “Xpointer”and “Xlink”.

2.Consequently, the DTD is quite large. Therefore the DTD should have a modular structure, as for example the DTDs of the Text Encoding initiativ (Burnard and Sperrberg-McQueen 1992).

3.Perspecta is just one of several programs which allow to visualize the connections between word. Other examples include *thinkmap* of the company plumbDesign or *The Brain* of Natrifical. Furthermore similar results can be achived with modelling languages like java3-d or *VRML.*

4. The price of this approach, however, is that the connection offered by the system might be not the link the user wants to pursue. If this is the case he or she should be able to switch to another strategy of knowledge exploration, e.g. the generation of excerpts about this time.