Journal of Universal Computer Science, vol. 8, no. 6 (2002), 581-590
submitted: 4/2/02, accepted: 3/5/02, appeared: 28/6/02 © *J.UCS*

Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources

Wolf-Fritz Riekert
(University of Applied Sciences Stuttgart – School of Media, Germany
http://v.hdm-stuttgart.de/~riekert/
riekert@hdm-stuttgart.de)

Abstract: There is an immense number of information resources on the Internet that can be utilized free of charge. So many knowledge workers try to make use of this information in their daily tasks. Nevertheless, it is very hard to find the relevant information in the Internet by using the full-text retrieval techniques which are offered by most existing search engines.

This paper demonstrates that Thesauri, which have been used in established online retrieval systems for a long time, also open up new methods for the automated search for information in the Internet. In addition, thesaurus-like structures known as Gazetteers allow handling geographical references of information resources in a very effective way. The knowledge represented in thesauri and gazetteers can be used to process a variety of thematic and geographical queries and to retrieve the information of interest from the Internet. Comfortable ways of specifying queries can be offered to the users, e.g., by navigating in a hierarchical tree of descriptors, by using synonymous, related or foreign-language terms rather than fixed elements of a controlled vocabulary, or by indicating a geographical region of interest on a cartographic map.

In addition to the general principles, examples of powerful query processors and advanced user interfaces are presented which demonstrate the effective usage of the knowledge stored in thesauri and gazetteers. The implemented solutions turn out to be considerably more comfortable than the "black box search" offered by most existing library catalogs and Internet search engines.
Key Words: Information Retrieval, Internet, Thesaurus, Gazetteer
Categories: H.3.3

1 Starting Point

The amount of information offered in the Internet is still rapidly increasing. On the one hand, this process is considered as a very positive one since it favors the building of an "Information Society". On the other hand, it hinders information users to keep the orientation within a plethora of information services. This problem, also referred to as "Lost in Hyperspace" syndrome, requires special systems and tools to effectively support the search for information in the Internet.

There are many search engines in the Internet which allow users to search for web pages by using a full-text retrieval facility. Full-text retrieval, however, processes the users' queries only in a textual way. There is no semantic interpretation of the search terms. A web page about an *inn* in *Graz* for instance, might not be found by such a search engine, if one searches for *accommodation* in *Styria*. On the one hand, a full-text index does not provide the terminological knowledge saying that the term *inn* is a synonym for the term *hotel* which in turn stands for a special kind of *accommoda-*

tion. On the other hand, a full-text search engine would lack the geographical knowledge that *Graz* is situated in the Austrian state of *Styria.*

So full-text retrieval is not sufficient for all application areas. Search criteria are required which are semantically deeper than ordinary full text matches. Towards this end, a special type of system also known as metainformation system has emerged. Such a system allows the indexing and retrieval of information by using criteria which are semantically deeper than the simple text patterns used in ordinary search engines. It turned out that three particular kinds of semantic descriptors are of major importance for a large number of information, namely *temporal references*, *thematic references*, and *spatial references* [Tochtermann et al. 1997]. This is also reflected by the established Dublin Core Metadata Standard, which provides the descriptors DC.SUBJECT for thematic references and DC.COVERAGE for spatial and temporal references [Dublin Core 2002].

The query in the example given above aims at information resources with the term "inn" as thematic reference and the term "Styria" as spatial reference. A temporal reference is not specified in the query; it could be given by the current date. Examples for metainformation systems which support these kinds of search criteria are the German Environmental Information Network (GEIN) [Bandholtz et al. 2000] and the Geographical Information System Environment (GISU) [Balzer and Nouhuys 1998] of the German Federal Environment Ministry.

2 Processing Thematic References with a Thesaurus

Keywords are the simplest way to specify thematic references. The keywords, however, should be taken from a *controlled vocabulary*, e.g., a *Thesaurus*. A thesaurus serves two purposes: On the one hand, it is a catalog of all terms that can be used for indexing information resources. On the other hand, a thesaurus is much more than a linear catalog since it represents terms as conceptual objects. Relationships link the terms in a thesaurus to each other thus forming a semantic network. There are basically three kinds of relationships in a thesaurus: (1) the "used-for"-relationship which allows finding terms in the controlled vocabulary starting from their synonyms, (2) a specialization hierarchy which reflects the relationship between broader and narrower terms, and (3) the linkage between related terms within the controlled vocabulary. Such a thesaurus is able to represent that the term "inn" can be used for the term "hotel" which, in turn, is a narrower concept than the more general concept "accommodation".

No doubt the first part of the search presented in the first section (for information with "inn" as thematic reference) can be supported by the knowledge represented in a thesaurus. In the sequel, system components are presented which make use of a thesaurus in order to support queries of the described kind.

3 A Thesaurus Navigator

Most existing metainformation systems resemble the well-known Online Public Access Catalogs (OPAC) that are available for many libraries. These systems implement

an interaction mode which is also referred to as *black box search*. In order to initiate a search, users have to fill a form with the desired search criteria. Unfortunately, there is no way to know in advance how many items will be hit by the entered criteria. Especially for casual users, it is often not clear how many restrictions have to be formulated in order to yield a reasonable result set and to avoid extreme situations such as zero or thousands of hits.

Therefore many users dislike search forms in metainformation systems and prefer browsing facilities that allow navigating in a hierarchical catalog of themes instead. The advantage of such a catalog is that users can immediately recognize if there is information available for a specific theme. In addition, navigation in hierarchical catalogs is a procedure which is very familiar to most computer users.

It is not easy, however, to maintain such a catalog manually. Automatic methods are required that continually update such a catalog. Such methods were developed in an R&D project conducted by the author at the Research Institute for Applied Knowledge Processing in Ulm [Riekert et al. 1999]. A thesaurus, namely the Environmental Thesaurus of the German Federal Environmental Agency [Batschi 1994], served as an information source for the generation of the catalog. The metadata were taken from the prototype for the German Environmental Information Network (GEIN prototype) [Tochtermann et al. 1997]. The methods developed were implemented in a prototypical software system.

Whenever there is a change in the metadata, the hierarchical catalog has to be regenerated by the system. This catalog basically consists of a "weeded" thesaurus that only contains the terms that are relevant for the existing metadata. For this purpose, the software system determines the subset of terms from the thesaurus which are actually being used as a thematic reference in some metadata record of an information resource. Recursively, the system adds all broader terms from the thesaurus hierarchy to the catalog until the top-level terms are being reached. It should be noted that the Environmental Thesaurus used is a poly-hierarchical thesaurus. The poly-hierarchical property means that any term in the thesaurus may have more than one broader term above it in the hierarchy and, thus, may appear more than once in the resulting catalog.

The presentation of and the navigation in the catalog is done with a tool also referred to as Thesaurus Navigator. The interface of the navigator is similar to Microsoft's Windows Explorer and consists of three display areas. The relevant terms from the thesaurus are shown as a folder hierarchy in the left display area. On a mouse click, narrower terms of relevance can be made visible as subfolders. The display area in the middle shows the names of information resources for which the selected term in the left area serves as a thematic reference. A mouse click on one of these names is sufficient to display the complete metadata in the right area. This metadata also comprises a hyperlink to the original information resource in case it is available in the Internet.

The navigator was implemented as a Java applet. Java Database Connectivity (JDBC) is being used to access the thesaurus and the metadata in the database. Therefore it is possible to invoke the navigator from any Java-enabled Internet browser. The installation of additional client software is not necessary.

The result of this project was a prototype that could demonstrate how navigation in a catalog can be used as an interaction mode in contrast to the traditional "black

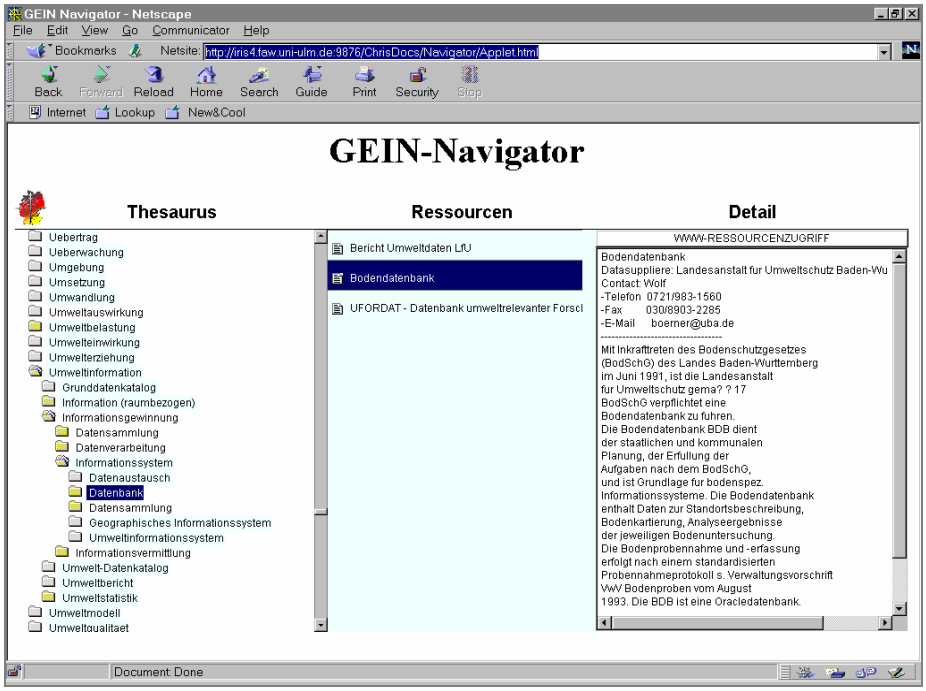


Figure 1: Navigation in a hierarchical catalog generated from the relevant terms in the environmental thesaurus.

box search". There was no formal evaluation of the prototype. Nevertheless, the reactions of the test users were mainly positive and supported the hypothesis that a large number of users would appreciate it very much if the interaction mode of navigation were provided by a catalog system at least as an additional option.

4 Enhancing a Search Engine with a Thesaurus

As it was shown in the previous sections, both metainformation systems and search engines support the search for information resources and the direct access to them via hyperlinks. Nevertheless both types of system differ from one another in the efforts necessary at the sides of the suppliers and users of information respectively:

- The maintenance of metainformation systems is a very time-consuming process for the information suppliers. The retrieval of information from a well-maintained metainformation system, however, is a very easy and comfortable process for the information users.
- Search engines impose nearly no work at all on the information suppliers since the indexing job is done automatically by a robot program. The users of the information, however, suffer from the fact that there is no semantic interpretation by the search engine, as already stated in the starting point section of this paper.

These observations were the motivation to develop a method to combine the advantages of both metainformation systems and search engines and to avoid most of the disadvantages of both kinds of system [Riekert et al. 1999]. The idea is to construct a thesaurus-based preprocessor that configures and reformulates the users' queries before they are transmitted to the search engine. This approach supports the retrieval of information in two ways:

- It is possible to search for information by navigating in the semantic network given by the thesaurus. During this navigation, the query can be built incrementally by continuously adding new visited nodes (i.e., terms) to the query.
- It is possible to reformulate queries in an intelligent way. The query can be augmented by adding other relevant terms which are narrower, broader, sibling, related or synonymous with respect to the entered criteria. With a multilingual thesaurus, it is also possible to translate the terms before the query is transmitted to the search engine.

Based on this approach, a prototype system was developed which again made use of the Environmental Thesaurus of the German Federal Environmental Agency [Fig. 2]. This thesaurus supports the two languages German and English. Alternatively, the General European Multilingual Environmental Thesaurus (GEMET) can be used which supports all major European languages but consists of a smaller number of terms. Both thesauri follow the same data model which complies with the international standards ISO 2788 and 5964. The implementation of the system is again based on Java and JDBC. The system uses Altavista [Altavista 2002] as a default search engine but it can be configured in a way that other search engines can be used as well. The Java program activates the search engine through the Hypertext Transfer Protocol (http) in the same way as a human user accesses it directly from an Internet browser. The result list produced by the search engine is directly transmitted to the user without further post-processing.

Investigations in this project showed, that search engines often fail if very specific information resources are being searched for. It could be demonstrated that the yield of useful information is improved considerably in many cases if the search encloses additional terms in the semantic neighborhood of the original search terms. Moreover, the optional translation of search terms often leads to a further improvement of the output. Depending on the setup of the various options [Fig. 3] the number of relevant results could be increased by factors between 2 and 100.

One could argue that this increase will make it more difficult to select the most relevant information and therefore the result list should be reduced rather than enlarged in order to remain manageable by the users. No doubt this argument is true in most cases if only one search criterion is being used. Queries with multiple search criteria, however, very often end up with the empty list. Here, a "softening" of the contributing search criteria is strongly desired in order to get results that match all criteria in a semantic rather than syntactic way.

In this project, it could be shown that it can be abstracted from the syntactic form of the queries. The users are supported by a semantic processing of their queries without imposing any additional work on the suppliers of the information. This makes the approach especially suited for the use in the Internet where the information users have practically no influence on the behavior of the information suppliers.

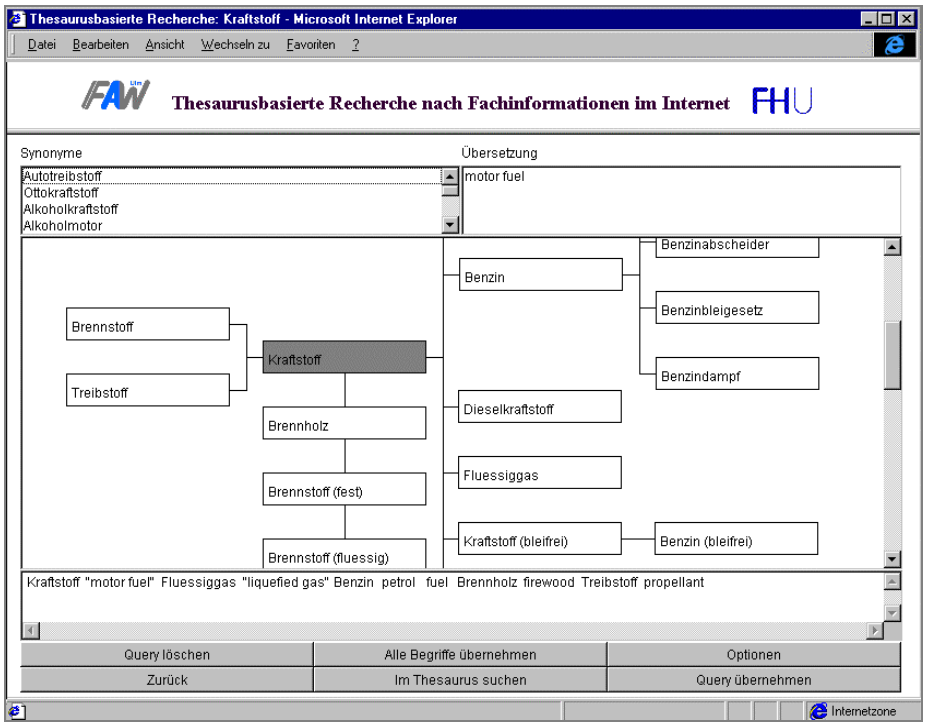
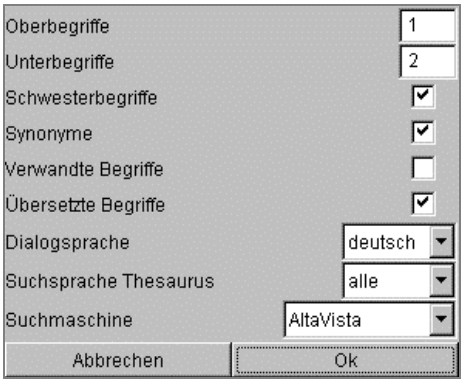


Figure 2: Thesaurus-based information retrieval in the internet: The user explores the thesaurus in the neighborhood of the German term "Kraftstoff" ("fuel"). Wider terms in distance 1 and narrower terms in distance 2 as well as synonymous and sibling terms are enclosed into the query as specified in the option sheet [Fig. 3]. The resulting query will be translated (currently only into the English language) and can be submitted to the search engine by a mouse click.



Distances to enclose wider and narrower terms into the query

Should sibling terms (i.e. terms with a common wider term), synonymous or related terms be enclosed?

Should terms be translated?

Language to enter the search terms

Language of the search terms

Search engine to be used

Control buttons (Abort, OK)

Figure 3: Search options: One level of wider terms, two levels of narrower terms, all sibling terms, synonyms and translated terms (but no related term) will be enclosed into this query. The search terms are to be specified in German, the search will be done in all languages by using Altavista as a search engine.

5 Processing Spatial References with a Gazetteer

The interpretation of the second part of the query in the beginning of this paper (for information related to "Styria") requires that information resources are spatially referenced. Spatial references can be specified in basically two ways: (1) textually, i.e., by indicating a geographic name, (2) geometrically, i.e., by specifying coordinates. The latter can be done either by defining numerical values or by a pointing on locations in an electronic map with the mouse. In principle, these two ways of indicating spatial references can be done at both indexing and retrieval time. The translation between both kinds of references can be done by using an information structure known as *Gazetteer*.

A gazetteer is a structured geographical index, in which each element has a geographical name and a geometry (e.g., a polygon described by coordinates such as longitude and latitude). By associating catalog items with gazetteer objects, catalog items can be searched for based on their geography. Particularly, in environmental settings, it is important that the catalog items have an n:m relationship to the gazetteer objects. For example, an environmental directive or law can have more than one geographical relationship (e.g., if it is valid in several geographical regions) and vice versa, in one geographical region several different environmental laws can be valid. Ideally, various layers of gazetteer objects including administration units (districts, states, and neighboring countries), water bodies (lakes, rivers, canals) or postal zones can be used for geographical references. Other than semantic relationships in a thesaurus, which are explicitly stored in a database, the relationships between gazetteer objects are represented implicitly. The gazetteer implicitly supports topological relationships such as *encloses*, *is enclosed by*, *is adjacent to,* and *overlaps*. By comparing the geometries (i.e., the coordinates) of the respective gazetteer objects, these relationships can be computed on demand whenever they are needed.

These implicit topological relationships between gazetteer objects can be exploited for both indexing of and searching for catalog items by using geographic descriptors (e.g., geometries or geographical names). That means particularly, that the system can infer that "Graz" is situated in "Styria". Unlike a thesaurus, where each new contribution must be evaluated on the basis of the existing content and where explicit relationships must be established with the existing contents of the thesaurus, this computational approach allows suppliers to keep the efforts required to enter new catalog items in the gazetteer as low as possible.

The geometry of a gazetteer object can be represented in different ways:

- The most advanced approach is to use vector data, i.e., polygonal approximations of the gazetteer objects. This approach allows the most precise representation of the geometries. Stock relational databases are not sufficient for this purpose; Geographic Information Systems (GIS) technology is required instead. An example for this approach is the geographic access system developed in the research project PADDLE (Personal Adaptable Digital Library Environment) [Tochtermann 2002].
- The simplest way is to represent the geometry as an enclosing rectangle. This occupies only four coordinates that can be easily stored in a relational database. This approach, however, has the disadvantage that the geometries are coarsened to a large extent thus lowering the precision of a geographic search considerably.

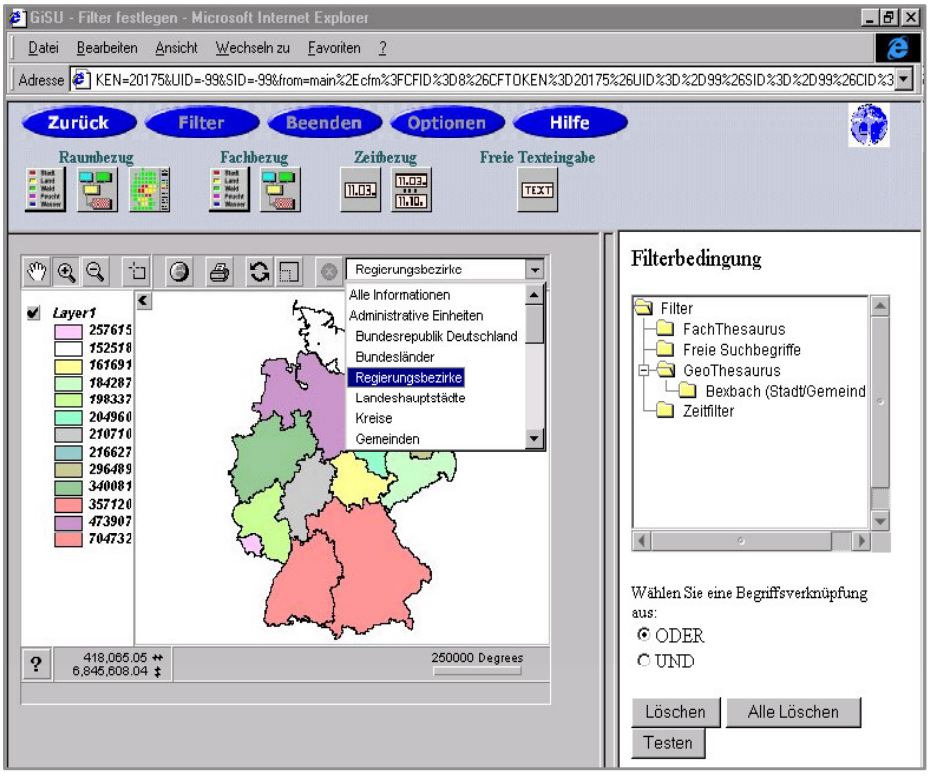


Figure 4: The spatial browser used in the GISU system

A compromise between both approaches had to be found for the GEIN prototype system. On the one hand, it was required to store all data in a commercial off-the-shelf database system. On the other hand, the enclosing rectangle approach could not fulfill the precision requirements. Therefore, a grid of rectangular raster cells was chosen as a spatial reference system. In this reference system, the geometry of a gazetteer object is approximated by the smallest collection of raster cells forming a complete coverage of the object. Since the location of a raster cell can be represented by a simple pair of coordinates, a table in a relational database can be used to represent the association of gazetteer objects with their related raster cells. By adjusting the size of the grid, the precision of the geometries could be adapted to the actual requirements. In this case, a grid size of 3×3km² turned out to be the best compromise between precision requirements and database performance.

A practical application of this concept is a gazetteer known as “Geothesaurus” which is being used in the systems GEIN and GISU of the German Federal Environmental Agency [Riekert and Treffler 2000]. Based on this Geothesaurus, a spatial browser has been developed for the GISU system by the German company Ernst Basler + Partner [Fig. 4]. With this browser, it is possible to retrieve information by indicating geographic names or geometries which are completely different from (but topologically equivalent with) the descriptors used at indexing time.

6 Conclusion and Outlook

The presented examples show that knowledge structures such as thesauri and gazetteers yield considerable advantages for the retrieval of information resources with metainformation systems and search engines. Gazetteers make improved spatial search facilities possible which will be necessary for the development of new, promising services. These services allow users to restrict their queries onto a particular geographic region such as the close neighborhood of their current position. Therefore the techniques described are also useful to support the new location-based services which are offered through an increasing number of mobile communication devices.

In addition, new attractive user interfaces can be built which allow the navigation through thematic references, the automated reformulation and translation of search criteria and the geographic search for information on a digital map.

The cost-efficiency of these solutions is considerably high. Other than the metainformation in digital catalogs and search engines, the information represented in a thesaurus or gazetteer is relatively stable, thus keeping the maintenance work small. In addition, this information is independent of concrete metainformation sets and can therefore be reused for a large number of different metainformation systems and search engines. This is particularly true for the Environmental Thesaurus and the Geothesaurus (gazetteer) of the German Federal Environmental Agency.

Therefore it is worth while investing into the construction of a thesaurus or gazetteer and the appropriate processing techniques and interfaces. New developments, as described in this paper, have a very positive impact on the effective use of the knowledge which is potentially available in the Internet.

References

[Altavista 2002] Altavista; http://www.altavista.com (2002).

[Balzer and Nouhuys 1998] Balzer, H. and van Nouhuys, J.: "GISU – Geographisches Informationssystem Umwelt im Umweltbundesamt"; In: Riekert, W.-F., Tochtermann, K. (eds.): Proc. Hypermedia im Umweltschutz. Ulm, Germany (1998). Metropolis-Verlag, Marburg (1998).

[Bandholtz 2000] Bandholtz, T., Bös, R., and Rüther, M.: "The German Environmental Information Network"; in: [Cremers and Greve 2000].

[Batschi 1994] Batschi, W.D.: "Environmental Thesaurus and Classification of the Umweltbundesamt (German Federal Environmental Agency) Berlin"; in: Stancikova, P., Dahlberg, I. (eds.): Environmental Knowledge Organisation and Information Management; Proceedings, Bratislava, Slovakia (1994). INDEKS Verlag, Frankfurt/Main (1994).

[Cremers and Greve 2000] Cremers, A.B. and Greve, B. (eds.): "Computer Science for Environmental Protection"; 12th Symposium, Proceedings, Bonn, Germany (2000); Metropolis Verlag, Marburg (2000).

[Dublin Core 2002] Dublin Core Metadata Initiative; http://purl.oclc.org/dc/ (2002).

[Riekert et al. 1999] Riekert, W.-F., Fuchs, Ch., and Klingler, G. (1999): "Erschließung von Fachinformationen im Internet mit Hilfe von Thesauri und Gazetteers"; in: Dade, C. and Schulz, B. (eds.): Management von Umweltinformationen in vernetzten Umgebungen; Proceedings, Nürnberg (1999); Metropolis-Verlag, Marburg (1999).

[Riekert and Treffler 2000] Riekert, W.-F. and Treffler, P.: "Georeferenzierung als Mittel zur Erschließung von Fachinformationen in Internet und Intranet"; in: [Cremers and Greve 2000].

[Tochtermann et al. 1997] Tochtermann, K., Riekert, W.-F., Wiest, G., Seggelke, J., and Mohaupt-Jahr, B.: "Using Semantic, Geographical, and Temporal Relationships to Enhance Search and Retrieval in Digital Catalogs"; in: Peters, C., Thanos, C. (eds.): Research and Advanced Technology for Digital Libraries; Proceedings ECDL'97; Springer-Verlag, Berlin (1997).

[Klaus Tochtermann 2002] Tochtermann, K.: "Personalisierung im Kontext von digitalen Bibliotheken und Wissensmanagement"; eingereichte Habilitationsschrift an der Technisch Naturwissenschaftlichen Fakultät der TU Graz (2002).