Treatment of Semantic Heterogeneity using Meta-Data Extraction and Query Translation

Robert Strötgen
Social Science Information Centre, Bonn

Summary

The project CARMEN[1] ("Content Analysis, Retrieval and Metadata: Effective Networking") aimed – among other goals – at improving the expansion of searches in bibliographic databases into Internet searches. We pursued a set of different approaches to the treatment of semantic heterogeneity (meta-data extraction, query translation using statistic relations and cross-concordances). This paper describes the concepts and implementation of these approaches and the evaluation of the impact for the retrieval result.

1 Treatment of Semantic Heterogeneity

Nowadays, users of information services are faced with highly decentralised, heterogeneous document sources with different kinds of subject indexing. Semantic heterogeneity occurs e.g. when resources using different systems for content description are searched by using a single query system. It is much harder to deal with this semantic heterogeneity than the technological one. Standardization efforts such as the Dublin Core Metadata Initiative (DC) are a useful precondition for comprehensive search processes, but they assume a hierarchical model of cooperation, accepted by all players.

Because of the diverse interests of the different partners, such a strict model can hardly be realised. Projects should consider even stronger differences in document creation, indexing and distribution with increasing „anarchic tendencies" (cf. Krause 1996, Krause/Marx 2000). To solve this problem, or at least to moderate it, we introduce a system consisting of an automatic meta-data generator and a couple of transformation modules between different document description languages. These special agents are able to map between different thesauri and classifications.

The first step in handling semantic heterogeneity should be the attempt to enrich the semantic information about documents, i.e. to fill up the gaps in the documents meta-data automatically. Section describes a set of cascading deductive and heuristic extraction rules, which were developed in the project CARMEN for the domain of Social Sciences.

The mapping between different terminologies can be done by using intellectual, statistical or neural network transfer modules. Intellectual transfers use cross-concordances between different classification schemes or thesauri. Section describes the creation, storage and handling of such transfers.

Statistical transfer modules can be used to supplement or replace cross-concordances. They allow a statistical crosswalk between two different thesauri or even between a thesaurus and the terms of automatically indexed documents. The algorithm is based on the analysis of co-occurrence of terms within two sets of comparable documents. The main principles of this approach are discussed in section .

1 Funded by the German Federal Ministry of Education and Research in the context of the programme "Global Info", FKZ 08SFC08 3.

We used intellectually and statistically created semantic relations between terms for query translation (cf. section) and evaluated the impact of this translation on the query result (cf. section).

2 Meta-Data Extraction

2.1 Approach

The goal of extracting meta-data during the gathering process is to enrich poorly indexed documents with meta-data, e.g. author, title, keywords or abstract – while the other methods described in the following section run during the retrieval process. This meta-data should be available for retrieval along with certain intellectually added meta-data, but with a lower weight.

The actual algorithms for meta-data extraction depend on file formats, domain properties, site properties and style properties (layout). No stable and domain independent approach is known so far. Until conventions for creating HTML documents change and become standardized towards a semantic web only temporary and limited solutions can be found.

Relevant documents from the Mathematics exist mostly in PostScript format. These documents (thesis papers) stored in an unstructured file format contain some meta-data of high relevance, which are marked only by layout information (e.g. font size) or special keywords. By using and modifying some tools as prescript from the New Zealand Digital Library, PostScript documents have been transformed and analysed. Abstract, keywords and classification terms can be extracted from these documents with good success.

Internet documents from the Social Sciences are mostly structured HTML files, but they use html features mainly for layout reasons, not as mark-up for content. Meta tags are infrequently used, and their syntax is often not correct. Different institutions use different ways of creating their Internet documents and a large number of documents contain no information about author or institution at all. This makes extraction of meta-data very difficult.

Nevertheless we developed a set of heuristics for identifying some meta-data in this heterogeneous set of documents. Because operating on (frequently incorrect) HTML files is not very comfortable, we transformed the documents into (corrected) XHTML and implemented our heuristics on these XML trees using XPath queries. (cf. Strötgen & Kokkelink 2001)

2.2 Evaluation

The test corpus for the Social Sciences contains 3.661 HTML documents collected from Web servers from different research and education institutions. The documents are of different types (e.g. university calendars, project descriptions, conference proceedings, bibliographies). We analysed these documents and used them as basis for the creation and improvement of the extraction heuristics.

From these documents 96% contain a correctly coded title; 17.7% of the rest contain an incorrectly marked title, the other documents contain no title at all. Only 25.5% contain keywords, all of them are marked properly. Not more than 21% contain a correctly coded abstract, 39.4% of the rest contain a differently marked abstract. This survey is the base for the evaluation of the meta-data extraction. Of course meta-data can only be extracted, if it is present in the document at all – we did try to implement automatic classification or automatic abstracting.

For the evaluation of the extracted meta-data we created a random sample containing every tenth document (360). We intellectually rated the relevance and correctness of the extracted data in four grades (accurate and complete; accurate in detail, but not complete or inaccurate parts; not accurate; not rateable).

We found that 80% of the extracted titles are of medium or high quality; almost 100% of the extracted keywords are of high quality; and about 90% of the extracted abstracts are of medium or high quality. (cf. Binder et al. 2002)

3 Query Translations using Semantic Relations

Intellectual and statistical semantic relations between terms or notations expand or modify the query during retrieval. The translation of the user's query, which was formulated for one of the document collections, leads to specific queries for each target document collection considering the different systems for content analysis.

3.1 Intellectual Transfer Relations (Cross-Concordances)

Intellectual transfers use cross-concordances between different controlled documentation languages like thesauri and classifications. These languages have a limited number of indexing terms or classes used to describe document subjects. A thesaurus is a natural language based documentation language with different kinds of vocabulary control and terminological control (i.e. synonym control and homonym control). Every thesaurus term is unique. Relations like equivalence, hierarchy and association are defined between the descriptors (cf. Burkart 1997). A classification is usually an artificial (alphabetical, numeric or alphanumerical) documentation language. A classification has classes or notions, which are systematically ordered with hierarchical relations, the classification structure. The class and its concept have a verbal class description (cf. Manecke 1997).

To build cross-concordances documentary and professional experts created semantic relations between thesaurus terms or classes with similar meanings.

Different kinds of inter-thesaurus or inter-classification relations are defined: "exact equivalence", "inexact equivalence", "narrower equivalence" and "broader equivalence". These relations can be annotated with weight information ("high", "medium", "low") reflecting the relevance of the relations for retrieval quality.

In the project CARMEN cross-concordances are provided between universal or special classifications and thesauri in the domains involved (mathematics, physics, social sciences). These cross-concordances allow safe transfers between different documentation languages and the document sets using them.

Problems may arise if there are insufficient resources (time, money, domain experts) to create and maintain such cross-concordances; furthermore not all documents – particularly Internet documents – are indexed with a controlled vocabulary. Therefore additional approaches like statistical transfers based on the analysis of co-occurrence of terms (see section) are necessary.

The software tool SIS-TMS[2] proved useable for creation of cross-concordances between different thesauri. CarmenX[3] has proved to be equally useful for creating relations between different classifications.

3.2 Statistical Transfer Relations

Quantitative statistical methods offer a general, automatic way to create transfer relations on the basis of actual bibliographic data. Co-occurrence analysis is one of those methods. It takes advantage of the fact that the content analysis from two different libraries on a single document held in both collections will represent the same semantic content in different content analysis systems. The terms from content analysis system A which occur together with terms from con-

2 http://www.ics.forth.gr/proj/isst/Systems/sis-tms.html
3 http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en

tent analysis system B can be computed. The assumption is that the terms from A have (nearly) the same semantics as the related ones from B, and thus the relation can be used as a transfer relation. The prerequisite for such computations is a parallel corpus, where each document is indexed by two different content analysis systems.

The classical parallel corpus is based on two different sets of documents, e.g. two different library catalogues. Each is indexed with a specific thesaurus/classification. To be able to create co-occurrence relations between the terms of these thesauri, the indexations of the documents have to be brought into relation. This is done by finding identical (or at least equivalent) documents in both catalogues. Considering print media, the problem of identity can be solved quite easy. An identical ISBN in combination with at least one identical Author should be a sufficient criterion for the identity of two documents. But the situation is worse, if the underlying data sets are not bibliographic ones, e.g. if text data should be combined with fact data or if Internet texts are considered.

In dealing with the World Wide Web we are concerned with many Web pages that are not indexed by a specific (given) thesaurus or classification system. The only terms we can rely on are the full-text terms supplied by a full-text indexing machine.

Taking into account that a user might start his search with controlled vocabulary obtained from a thesaurus, relevant Internet documents should be retrieved as well as well-indexed documents stored in a database. In order to facilitate this, we have to realize a transfer from classification terms, thesaurus terms respectively, to full-text terms, and vice versa. As long as we cannot fall back to any standards of classifying Internet documents, we have to use a weaker strategy of combining keywords and full-text terms. Note that intellectual indexing would result in enormous costs. This weaker strategy is simulating parallel corpora for supplying semantic relations between keywords and Internet full-text terms.

First of all, in order to provide a simulated parallel corpus, we have to simulate intellectual keyword indexing on the basis of a given thesaurus. Simulating intellectual indexing implies that a method is used that produces *vague* keyword-document-relationships, i.e. unlike intellectual indexing, where each assignment is (usually) un-weighted (weighted 1 respectively) simulated keyword-document-ties are weighted on a [0,1]-scale. This yields a situation as indicated in Figure 1: Unlike the situation in public databases (like the German social science literature database SOLIS), where we have exact assignments of keywords and documents, we produce vague keyword indexing as well as vague full-term indexing.

Parallel corpora simulation via vague keyword and full-text term assignments is described for the CARMEN test corpus. The CARMEN test corpus is a collection of about 6.000 social science Internet documents that have no keyword assignments.

For the assignment of controlled vocabulary (keywords) to non-classified Internet documents a given thesaurus is used, i.e., in the case of the CARMEN corpus, the thesaurus for Social Sciences (IZ 1999). As a basic method assigning thesaurus keywords to Internet documents we consider each single keyword of the thesaurus as a query that is “fired” against a full-text indexing and retrieval machine maintaining the CARMEN corpus. The full-text retrieval engine Fulcrum SearchServer 3.7 is used to provide ranking for values the documents in the corpus according to their similarity to the query vector. Each document in the result set that has a ranking value greater then a certain minimum threshold is then indexed by the keyword requested. The ranking values supplied by Fulcrum are considered the weights for the keyword assignments. This basic method has been improved to consider the relevance of a keyword for the document retrieved (cf. Hellweg et al. 2001).

Full-text terms are obtained by tokenising the full-text of an Internet document, eliminating stop words, and stemming the remaining terms using a Porter stemming algorithm for German. For weighting the terms the *inverse document frequency* (cf. Salton 1987) is used. The full-text is then indexed with full-text terms having a weight greater than a certain minimum threshold.

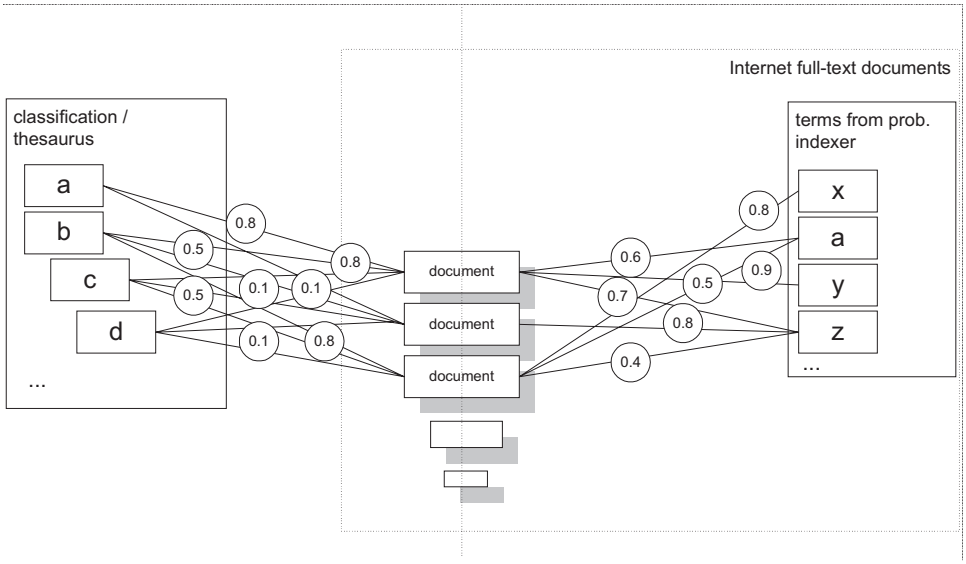

Figure 1: Parallel Corpus Simulation with vague keywords and full-text terms

In the context of the project ELVIRA, a tool for generating statistical correlation relations based on parallel corpora was implemented. JESTER (the Java Environment for Statistical Transformations) is a general workbench that allows for the interactive selection of parameters for optimising the transfer relation between a pair of classification systems. JESTER also employs a number of heuristics for the elimination of systematic errors, introduced by the simulation of an actual parallel corpus as described before.

In particular, the graphical representation of the term-document frequencies permits the eliminations of documents and/or terms from the following steps, based on their occurrence. In the case of a simulation of a parallel corpus, some documents got too many terms assigned. This happens, when the probabilistic search engine erroneously returns the same document on almost every query because some domain specific phrase appears verbatim.

3.3 Query Translation

Once the transfer relations have been realized, the question remains how to incorporate them into information retrieval systems. As a query term manipulating methodology they have to be placed somewhere between the user interface and the databases. Transfer relations, or – to be precise – transfer modules, become necessary, if data sets have to be combined, which are indexed by different content analysis systems. Usually those data sets reside in different databases. In a classical coordination layer the user query is simply distributed – unchanged – to the different databases and the results are combined to an overall result set. Most of the meta search engines in the WWW work this way.

But this procedure is not applicable to data with heterogeneous indexing systems. To send one and the same query to all databases would lead to a lot of zero results in the connected databases. This is due to the fact that e.g. a queried classification is available in only one database, but not in the others. At this point the transfer relations come into play. Through their ability to transform controlled vocabulary, they are able to adapt the users query to the different requirements of the

database. Therefore the query the user has issued will be transformed into different queries fitting the different content description systems (cf. Figure 2).

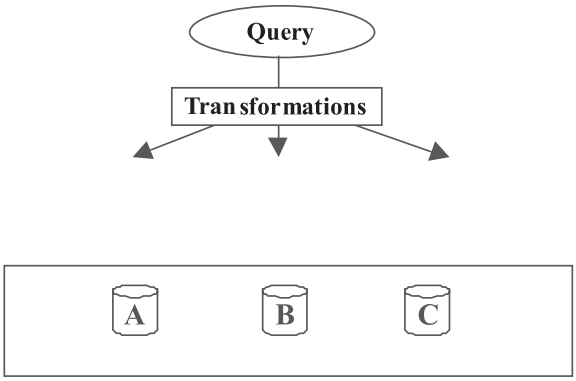

Figure2 : Query Translation Process

During the actual transformation the relevant part of the users query (e.g. the classification terms from system *A*) is separated from the other parts. The terms of this separated part act as the input for the different transformation modules. After the transformation, the resulting output forms the new, transformed query part. This new part consists of terms from system *B*, and is combined with the rest of the users query (e.g. author/title query) to form the new, transformed query. Afterwards the query is sent to the corresponding database.

This procedure follows the so-called "two-step" model (cf. Krause 2000) developed by the Social Science Information Centre (IZ) in the context of different projects.[4]

For CARMEN this approach was implemented in the project's architecture. The retrieval system *HyRex*[5] is part of a package developed at the University of Dortmund (cf. Fuhr et al. 2000). This search engine uses transfer services running remotely on servers at the Social Science Information Centre. Relevant parts of the complete query (e.g. author is no transferable query type) are sent to the transfer service as XIRQL (XML *Information Retrieval* Query Language) statements by http request and answered with a new transferred XIRQL statement that represents the transferred query (cf. Figure 3).

4 ELVIRA, CARMEN, ViBSoz and ETB (cf. Hellweg et al. 2001).
5 http://ls6-www.informatik.uni-dortmund.de/~goevert/HyREX.html

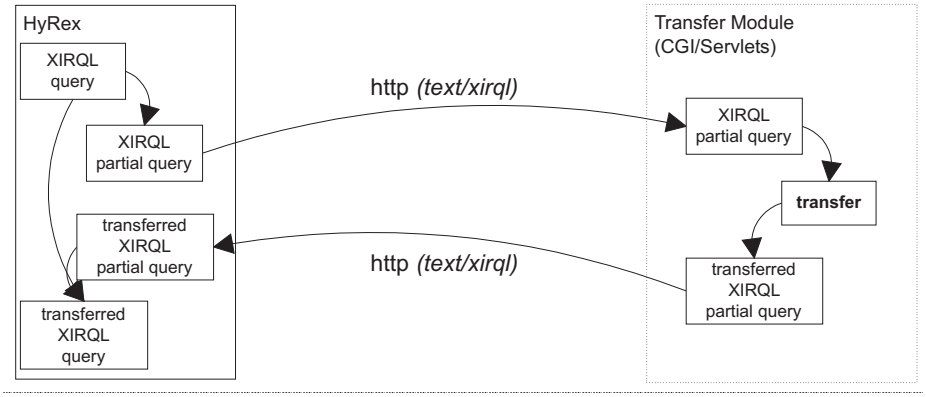

Figure 3: Query Transfer Architecture

3.4 Evaluation

In order to evaluate the impact of query translation with statistically created transfer relation we performed retrieval tests. We indexed about 10.000 HTML documents using Fulcrum Search Server (HyRex was not available for the retrieval tests in time). Using Fulcrum we did not make use of the weight information attached to semantic relations; this feature is implemented in HyRex, but the effect is lost for the test.

Our scenario consisted of a search starting from the bibliographic database SOLIS using the "Thesaurus Sozialwissenschaften" as documentation language for query formulation. This query was supposed to be expanded to Internet documents. For this purpose the query is translated from the controlled thesaurus term to free terms. In this special case no translation of one controlled language to another leads to a replacement of terms; the uncontrolled free terms as translation target allow the addition of semantically related terms.

We executed the search in two ways: We sent both the original SOLIS query and the translated (expanded) query to the retrieval engine and compared the results. We have not been able to perform representative tests but exemplary spot tests. For each of three domains from the Social Sciences (women studies, migration, sociology of industry) we carried out two searches. Two of them are described exemplarily:

One query used the SOLIS keyword "Dominanz" ("dominance") and returned 16 relevant documents from the test corpus. Using the query translation 9 additional terms were found: "Messen", "Mongolei", "Nichtregierungsorganisation", "Flugzeug", "Datenaustausch", "Kommunikationsraum", "Kommunikationstechnologie", "Medienpädagogik", "Wüste").

You have to be very careful interpreting the statistically created relations in a semantic way. In this example the new terms "Mongolei" ("Mongolia") and "Wüste" ("desert") result of documents describing an excursion of a women activist group to China with a stop in the Mongolian desert. One should not deduce a "dominance problem" in Mongolia or in desert regions from this relation, but in other cases you will find semantic relations not complying semantic equivalence but a problem field.

With this expanded query 14 new documents were found; 7 of them were relevant (50%, a gain of 44%). The precision of this search is 77%. In this case without too much noise a significant number of new documents could be reached with the query translation.

Another, less successful example: A query using the keyword “Leiharbeit” (“temporary employment”) returned 10 relevant documents. The query translation added 3 new terms: “Arbeitsphysiologie”, “Organisationsmodell”, “Risikoabschätzung” and produced a result of 10 new documents, but only 2 of them were relevant (20%, a gain of 20%). The precision of this search is 60%. In this example the translated query results in very little gain of new relevant documents but 80% noise.

Summerising all examples we can state that we always found new relevant documents using the translated query compared with the original query. The precision of the additionally found documents ranges between 13% and 55%. Without being already able to find systematic conditions we find rather successful and weaker query results.

4 Outlook

The modules for meta-data extraction proved to be satisfactory. Meta-data was extracted tolerably from Web documents. They have been integrated in the gathering system (“CARA”) and can be used for other projects. It seems promising to transfer the heuristics for HTML documents to other domains than the Social Sciences. Probably the weighting component will need some adjustment.

The modules and heuristics are in general functioning; some improvement is conceivable by tuning the heuristics. Because of the transient Web standards and the fast changes in Web style very high effort for maintenance seems necessary to keep the heuristics up to date. It seems questionable if this effort can be raised.

The query transfer modules using statistically created semantic relations proved able to improve the query result in retrieval test. New relevant documents were found using the transfer from a thesaurus to free terms for an Internet search, but some queries produce more noise then useful documents.

Some aspects remain unanswered and require more research and tests.

How can the document corpus, used for computing the semantic relations, be improved; e.g. what kind of documents create bad artefacts or which properties does a corpus need to be representive? Probably the statistical methods need some refinement.

In the project’s context intellectually created cross-concordances have been created and evaluated separately. The tested transfer modules can handle both kinds of semantic relations, and both should be compared directly. Also methods of combining both ways should be implemented and evaluated.

Of course the user interaction remains an important topic. By now the transfer process is a black box for the user. An user interface is needed that allows the user to understand and to influence this process and its parameters without handling incomprehensible numbers like statistical thresholds. The outcome of an interactive retrieval using transfer modules must be evaluated with real user tests.

An output of the project are services and software modules for query translation, which are offered to interested users. We already integrated them into existing services like “ViBSoz” (Virtuelle Fachbibliothek Sozialwissenschaften); other new services like “ETB” (European Schools Treasury Browser) and “Informationsverbund Bildung – Sozialwissenschaften – Psychologie” will follow.

5 References

Binder, G..; Marx, J.; Mutschke, P.; Strötgen, R.; Plümer, J.; Kokkelink, S. (2002): Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht Nr. 24).

Burkart, M. (1997): Thesaurus. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Eds.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Dokumentationsarbeit. München, p. 160 - 179.

Fuhr, N.; Großjohann, K.; Kokkelink, S. (2000): CAP7: Searching and Browsing in Distributed Document Collections. In: Borbinha, José; Baker, Thomas (Eds.): Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000; Lisbon, Portugal, September 18-20, 2000; Proceedings. Berlin: Springer 2000. (Lecture notes in computer science ; Vol. 1923). p.364 - 367.

Hellweg, H. (2002): Einsatz von statistisch erstellten Transferbeziehungen zur Anfrage-Transformation in ELVIRA. In: Krause, Jürgen; Stempfhuber, Maximilian (Hrsg.). Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA. (Forschungsberichte des IZ Sozialwissenschaften Band 4) (to appear 2002).

Hellweg, H.; Krause, J.; Mandl, T.; Marx, J.; Müller, M.; Mutschke, P.; Strötgen, R. (2001): Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht Nr. 23).

Krause, J. (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung; Schalenmodell. Bonn (IZ-Arbeitsbericht Nr. 6).

Krause, J. (2001): Virtual libraries, library content analysis, metadata and the remaining heterogeneity. In: ICADL 2000: Challenging to Knowledge Exploring for New Millennium: the Proceedings of the 3rd International Conference of Asian Digital Library and the 3rd Conference on Digital Libraries, Korea, December 6 - 8, 2000, Seoul, p. 209 - 214.

Krause, J.; Marx, J. (2000): Vocabulary Switching and Automatic Metadata Extraction or How to Get Useful Information from a Digital Library. In: Information Seeking, Searching and Querying in Digital Libraries. Proceedings of the First DELOS Net-work of Excellence Workshop. Zurich, Switzerland, December 11-12, 2000. Zurich. p. 133 - 134.

Manecke, H-J. (1997): Klassifikation. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Eds.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Dokumentationsarbeit. München, p. 141 - 159.

Salton, G. (1987): Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg – New York.

Strötgen, R.; Kokkelink, S. (2001): Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In: Schmidt, Ralph (Ed.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarkt; 23. Online-Tagung der DGI und 53. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Frankfurt am Main, 8. bis 10. Mai 2001; Proceedings. Frankfurt am Main: DGI 2001. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 4), p. 56 - 66.

6 Contact Information

Robert Strötgen
Informationszentrum Sozialwissenschaften
Lennéstr. 30
53113 Bonn
Germany

e-mail: stroetgen@bonn.iz-soz.de
www.gesis.org/iz