DEALING WITH SEMANTIC HETEROGENEITY AT THE QUERY PROCESSING LEVEL (THE SOCIAL SCIENCE VIRTUAL LIBRARY PROJECT - VIBSOZ)

Dr. Jutta Marx
Matthias N.O. Mueller
Social Science Information Centre
Lennestr. 30
D-53113 Bonn
+49/228/2281-0
jm@bonn.iz-soz.de
mr@bonn.iz-soz.de

ABSTRACT

The Social Science Virtual Library Project (funded by the 'Deutsche Forschungsgemeinschaft, DFG') aims at presenting an integrated view to distributed, heterogeneous data of German social science literature. The main emphasis has been put on solving problems of access to such diverse document sets. As a prerequisite for higher services, an adequate system architecture has been implemented. The heterogeneity in content description systems will be solved by translation components, which realize a switching of vocabulary.

1. INTRODUCTION

The landscape of research information of the social sciences in Germany is showing a great diversity in terms of relevance to the subject, quality of content analysis and the database system used for storage and retrieval [Krause 2000]. There are some special libraries, some general libraries with large amounts of social science literature and one central bibliographic database (SOLIS). All of these do not only have different user interfaces but worse, they use different systems of content description like thesauri and classifications. The goal of the project is to give the user a central access point with a single user interface and an integrated view of the existing thesauri and classifications. So the central point is to give the user the ability to search different databases which use different vocabulary for content analysis with just one query and to present the overall result in a single, uniform result set. To reach this goal, three tasks have been fulfilled:

implementation of an architecture which is able to integrate different information systems

solving the heterogeneity in content description problem, regarding the simultaneous use of different thesauri and classifications

implementation of an user interface which is easy to use and able to cope with the problems arising from the distribution of the system

2. ARCHITECTURE

The architecture is based on a three layered client/server model (cf. Figure 1). The first layer consists of different user/system interfaces. We developed a java client tailored to the special needs of our system. It is complemented by a Z39.50 server interface to allow the access with standard bibliography tools or the integration into other library systems. The second layer is made up of a central broker, which is able to handle the incoming user requests. It will process the user queries to fit the different semantics and structures of the databases connected (see sections 3 and 4), and will integrate the results returned.

The communication between the second and third layer (broker – databases) is handled via the popular Z39.50 protocol. So it is possible – in general – to integrate every Z39.50 server. This opens the possibility to integrate various other libraries, e.g. the local university library.

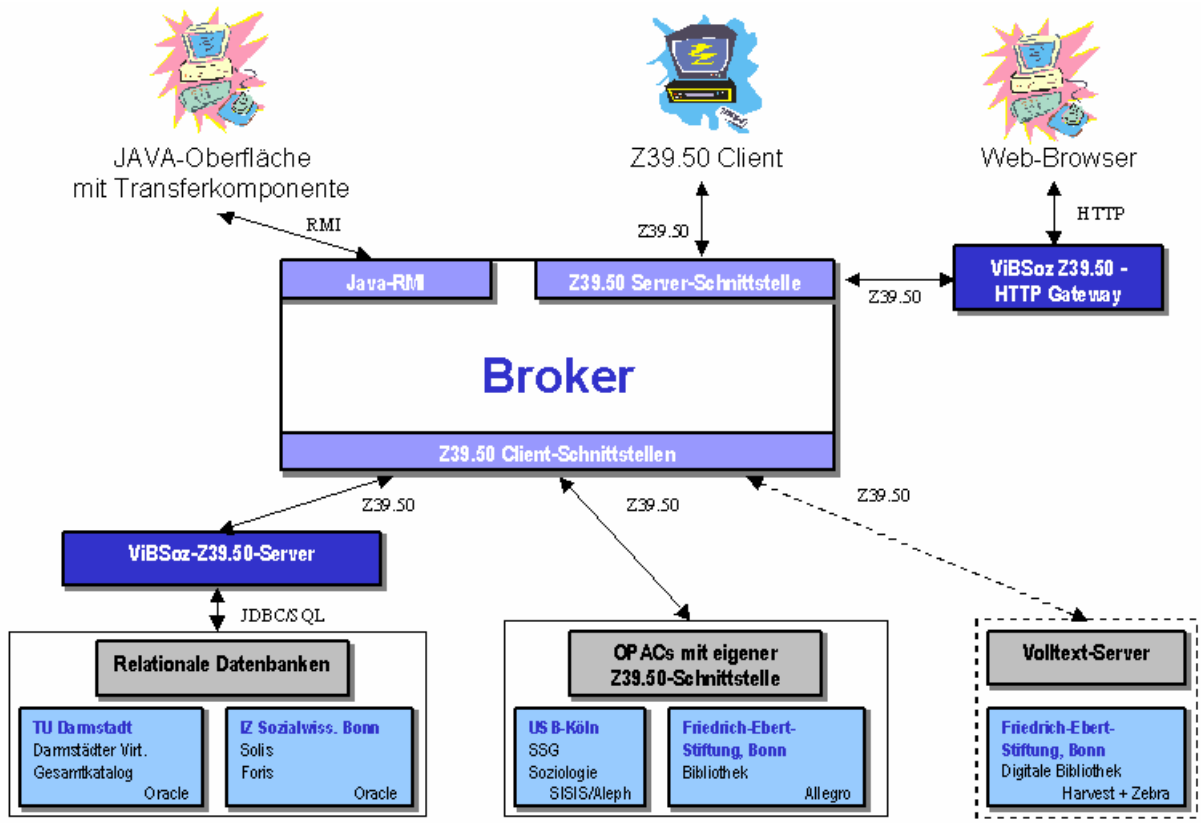

Figure 1: The ViBSoz-Architecture

3. SEMANTIC HETEROGENEITY

3.1 General Considerations

From the use of different thesauri and different classifications in one virtual data pool the problem arises that the user cannot use just one single thesaurus or classification to query all of the database. This problem space we call semantic heterogeneity. As a solution to it the Social Science Virtual Library provides translation components for query processing. These components provide a mapping between different thesauri or classifications, and thus enable the user to use only one of them to query all of the databases connected to the system.

Let's consider the German compound 'Jugendarbeitslosigkeit' (youth unemployment) as a simple example of such a translation. It is a composition of the two nouns 'Jugend' (youth) and 'Arbeitslosigkeit' (unemployment). The SWD (Subject Authority File of the German national library, Die Deutsche Bibliothek) uses this term in its composed form – so it is a precoordinated system – whereas the Thesaurus for the Social Sciences (edited by the IZ) prefers a combined form, consisting of the phrase 'Jugendlicher and Arbeitslosigkeit' – it is a postcoordinated system. The resulting mapping between the different terms of the example than has to be:

'Jugendarbeitslosigkeit' → 'Jugendlicher' AND 'Arbeitslosigkeit'.

Besides the difference of general and specialised thesauri, this difference between post- and precoordination accounts for a lot of the relations found.

To realize such mappings, three different methods are in use [Krause/Marx 2000]:

first, intellectual cross concordances similar to those used in retro cataloging of existing data with another classification,

second, statistical translation relations based on co-occurrence, and

third, we made experiments with neural networks [Mandl 2000].

In this paper, we will look at the statistical translation relations only.

3.2 Statistical Translation Relations

In contrary to intellectual relations statistical ones are not based on human knowledge about the problem space, but are extracted from an existing corpus of documents by mathematical methods. So these relations are more quantitative than qualitative relations.

To generate statistical translation relations a parallel corpus had been constructed. A 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 (cf. Figure 2). This is done by finding identical (or at least equivalent) documents in both catalogs. 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.

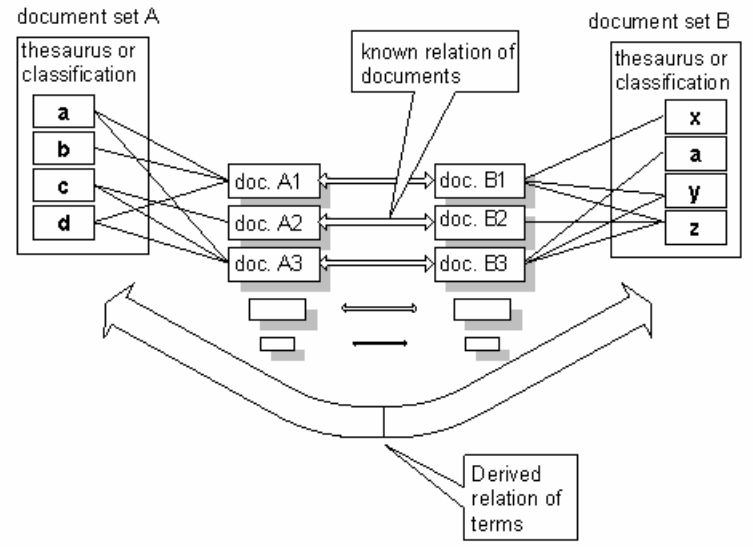

Figure 2: Parallel corpus

After the creation of the parallel corpus, the indexation terms of document Ax from data set A can be brought into relation with the indexation terms of the same document Bx from data set B. Lets consider the following document as an example:

'Gysi, Jutta: Familienleben in der DDR, zum Alltag von Familien mit Kindern, Akademie Verlag Berlin, 1989, ISBN 3-05-000771-0.'

It has been indexed by the library of the University of Cologne with the terms

'Deutschland <DDR>' (Germany <GDR>) and *'Familie'* (family)

from the SWD. Whereas the same document has been indexed by the IZ with the terms

'Arbeitsteilung' (division of labor), *'Ehe'* (marriage), *'Familie'* (family), *'DDR'* (GDR) and *'Partnerschaft'* (partnership)

from the Thesaurus for the Social Sciences.

Now the different terms can be brought into relation. This is done by relating every term from indexation A to every term of indexation B, e.g. the SWD term *'Deutschland <DDR>'* will be related to the social science thesaurus term *'DDR'*. Thus, this example document will result in ten relations (two SWD terms multiplied by five social science thesaurus terms). Of course not all of these relations make sense. The few really useful ones will be filtered out by a co-occurrence analysis. This method is similar to those used in some term expansion systems (e.g. [Biebricher et al. 1988]). The result of this process is a term-term-matrix of thesaurus A and thesaurus B with a statistical weight of the closeness of those terms (probability). Table 1 is showing some relations out of the current mapping. A more detailed description of the process and the tools used will be given in [Hellweg et al. 2001].

Term A	Term B	Probability
Habermas, Juergen	Habermas, J	0.977
Gewalttaetigkeit	Gewalt	0.883
Bevoelkerungsentwicklung	Bevoelkerung	0.770
	Entwicklung	0.688

Table 1: Examples of translation relations

3.3 Integration Into the System

Usually such translation mechanisms are realised at the database level. The data is simply enriched by another classification or thesaurus. But this procedure is quite inflexible: Every new database which should be integrated into the system would have to be enriched by at last one (general or specialised) thesaurus and classification. So we decided not to integrate the translations at the database level, but at the query processing level. The databases stay untouched – instead the query is manipulated to fit the data.

The process is as follows: The user can formulate his query using the java client developed, stating which thesaurus and which classification he has used for his formulation. This 'original' query is then sent to the broker along with the vocabulary information. The system is now able to translate this query into many other queries (cf. Figure 3) fitting the different vocabulary needs. Afterwards those queries are manipulated by e.g. adding a subtitle query field, and sent to the according database. So for each database connected to the library a separate and specialised query is generated. The results returned from each database are different in terms of format (MAB/MARC/XML), content, and sorting. Some databases supply only the minimum number of document fields, other are quiet extensive (e.g. contain an abstract). Some servers are able to sort the result set, others are not. So the combination of the different results to a uniform result sets is quite complex. First, all resulting documents are converted to an internal XML format. This format is then converted to the different output formats provided by the system.

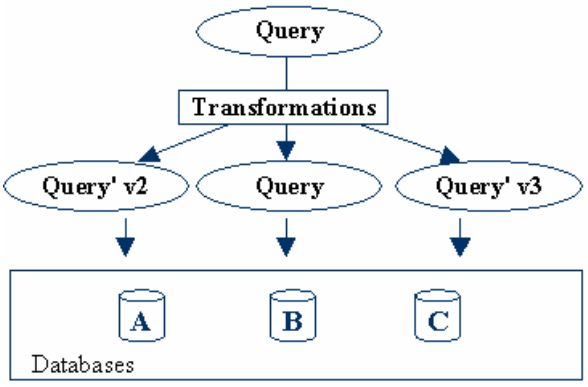

Figure 3: Query manipulation in ViBSoz

4. OUTLOOK TO PHASE II

A first version of the system has already been implemented. A report on the results of recall and precision tests will be given together with the paper presentation.

During the second phase of the project, starting in may 2001, are emphasized three new areas. These are the integration of new conventional libraries, the integration of world wide web documents / digital libraries and multilingual retrieval. Therefore an English language user interface will be developed.

5. REFERENCES

Biebricher, Peter; Fuhr, Norbert; Lustig, Gerhardt; Schwantner, Michael; Knorz, Gerhardt (1988): "The Automatic Indexing System AIR/PHYS - From Research to Application." In *11th International Conference on Research & Development in Information Retrieval*, Ed. Chiaramella, Yves. Grenoble, France: ACM Press.

Hellweg, Heiko; Krause, Juergen; Mandl, Thomas; Marx, Jutta; Mueller, Matthias N.O.; Mutschke, Peter; Stroetgen, Robert (2001): *Treatment of Semantic Heterogeneity in Information Retrieval*. IZ-Arbeitsbericht 23. Bonn: InformationsZentrum Sozialwissenschaften.
(See http://www.gesis.org/en/publications/reports/iz_working_papers/index.htm)

Krause, Juergen (2000): "Virtual libraries, library content analysis, metadata and the remaining heterogeneity." In *3rd International Conference of Asian Digital Library Conference (ICADL2000)*. Seoul.

Krause, Juergen; Marx, Jutta (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. First DELOS Network of Excellence Workshop*. Zurich, Switzerland.

Mandl, Thomas (2000): *Einsatz neuronaler Netze als Transferkomponenten beim Retrieval in heterogenen Dokumentbestaenden*. IZ-Arbeitsbericht 20. Bonn: InformationsZentrum Sozialwissenschaften. 116 p.