The Social Science Virtual Library Project
Dealing with Semantic Heterogeneity at the Query Processing Level

Dr. Jutta Marx
Matthias N.O. Mueller

Social Science Information Centre
Lennestr. 30
D-53113 Bonn

jm|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 systems used [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 Social Science Virtual Library project (ViBSoz) is to give the user a central access point with a single user interface and an integrated view of the existing thesauri and classifications. To reach this goal, three tasks have been fulfilled:

- implementation of an architecture which is able to integrate different information systems
- solving the problem arising from the heterogeneity in content descriptions, especially from 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 of the system is based on a three layered client/server model (c.f. Figure 1). The top layer consists of different user/system interfaces. As the primary access point we developed a java client tailored to the special needs of our system. It is complemented by a Z39.50 server interface to allow 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 processes the user queries to fit the different semantic and structural needs of the databases connected, and integrates the results returned from those. The bottom layer consists of the different databases integrated into the system. 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.

So the semantic heterogeneity in terms of multi database systems is resolved by the encapsulation of the different databases through a standardised query protocol. The system itself uses a 'mediator' (broker) which supplies a pure 'virtual schema' (the standard document fields defined in Z39.50) to enable a global search. Any further problems arising from the database community's definition of 'semantic heterogeneity' (see e.g. [Hull 1997]) are not considered relevant to the project.

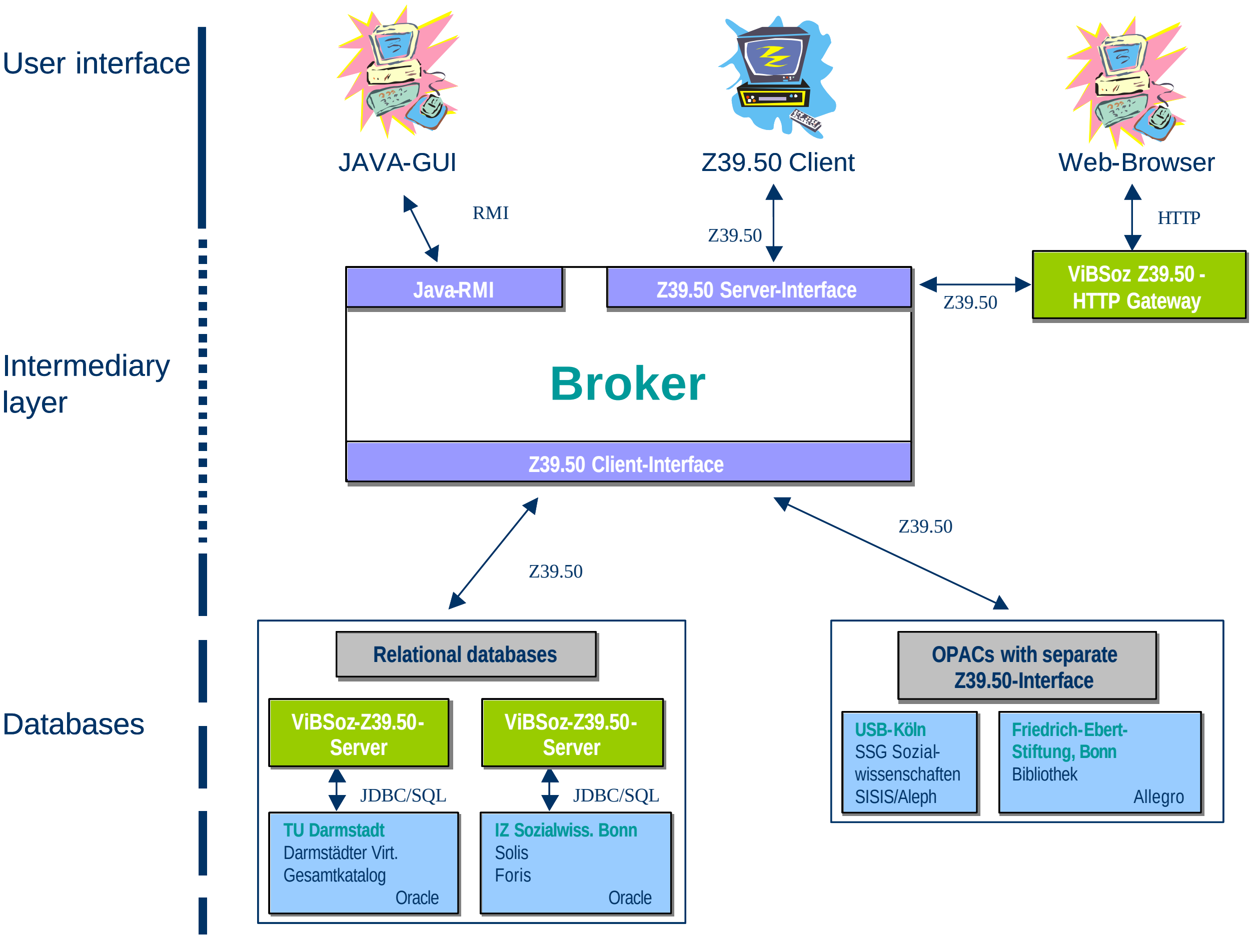

Figure 1: ViBSoz Architecture

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. So in contrary to the database community we are not concerned with the semantics of the data structures/schemas, but the semantic of the data itself.

As a solution to this heterogeneity in the data of the different systems the Social Science Virtual Library provides translation components for query processing. These components provide a mapping between different thesauri or classifications, and so relating the different expressions of the same semantic content. Thus it enables 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 Social Science Information Centre) 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 cataloguing of existing data with another classification,
- second, statistical translation relations based on co-occurrence, and
- third, we made experiments with neural networks.

This paper considers only statistical translation relations. For a more general description of translation relations as applied in the project see [Hellweg et al. 2001].

3.2 Statistical Translation Relations

Statistical translation relations are extracted from an existing corpus of documents by mathematical methods. Therefore those relations are not based on human knowledge about the problem space (like intellectual relations are), but reflect the actual use of the indexation terms in the underlying data. So they are more quantitative than qualitative relations.

To generate statistical translation relations a parallel corpus had been constructed. This corpus contains documents for which a content description (by thesaurus and/or classification terms) was provided by two different institutions (libraries or information centres). So to each document in the corpus terms of two thesauri and two classifications are assigned (c.f. Figure 2).

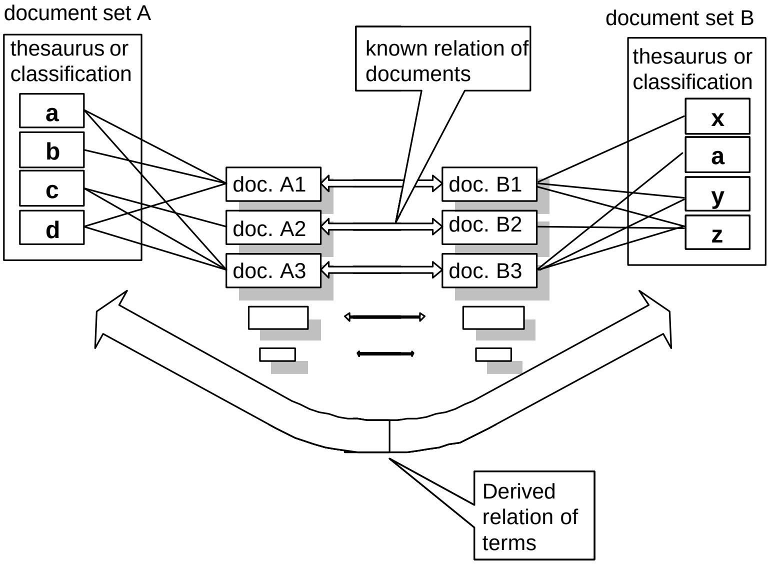
document set A
thesaurus or classification
a
b
c
d
known relation of documents
document set B
thesaurus or classification
x
a
y
z
doc. A1
doc. A2
doc. A3
doc. B1
doc. B2
doc. B3
Derived relation of terms

Figure 2: Parallel Corpus

After the creation, 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.’

This document 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 Social Science Information Centre with the terms

'Arbeitsteilung' (division of labour), *'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 is 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. 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 one classification. So we decided not to integrate the translations at the database level, but at the query processing level. The databases stay untouched but the query is manipulated to fit the data

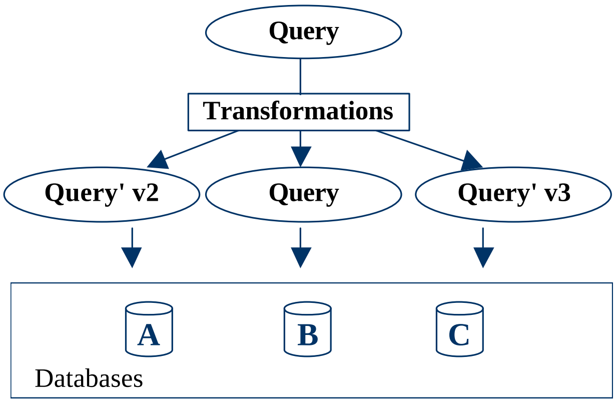

Figure 3: Query Manipulation in ViBSoz

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 fitting the different vocabulary needs (resolving semantic heterogeneity in a information science sense). Afterwards those queries are manipulated by e.g. adding a subtitle query field (resolving semantic heterogeneity in a multi database systems sense), and sent to the according database. So for each database connected to the library a separate and specialised query is generated. Afterwards, the results returned from each database are combined to a uniform result set and then are presented to the user.

A first version of the system has already been implemented. During the second phase of the project, starting in may 2001, the emphasis is on 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.

4. 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, Jürgen; Mandl, Thomas; Marx, Jutta; Müller, Matthias N.O.; Mutschke, Peter; Strötgen, Robert (2001): *Treatment of Semantic Heterogeneity in Information Retrieval*. IZ-Arbeitsbericht 23. Bonn: InformationsZentrum Sozialwissenschaften.

Hull, Richard (1997): "Managing Semantic Heterogeneity in Databases: A Theoretical Perspective." In *Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Tucson, Arizona: ACM.

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

Krause, Jürgen; 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.