The Domain-Specific Track at CLEF 2007

Vivien Petras, Stefan Baerisch, Maximillian Stempfhuber
GESIS Social Science Information Centre, Lennéstr. 30, 53113 Bonn, Germany
{vivien.petras | stefan.baerisch | max.stempfhuber@gesis.org}

Abstract

The domain-specific track uses test collections from the social science domain to test monolingual and cross-language retrieval in structured bibliographic databases. Special attention is given to the existence of controlled vocabularies for content description and their potential usefulness in retrieval. Test collections and topics are provided in German, English and Russian. This year, a new English test collection (from the CSA Sociological Abstracts database) was added. We present an overview of the CLEF domain-specific track including a description of the tasks, collections, topic preparation, and relevance assessments as well as contributions to the track. A summary of results is given. The track participants experimented with different retrieval models ranging from classic vector-space to probabilistic to language models. The controlled vocabularies were used for query expansion or as bilingual dictionaries for query translation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Evaluation, Controlled Vocabularies

1 Introduction

The CLEF domain-specific track evaluates mono- and cross-language information retrieval on structured scientific data. A point of emphasis in this track is research on leveraging the structure of data in collections (i.e. controlled vocabularies and other metadata) to improve search. In recent years, the focus of the domain-specific data collections was on bibliographic databases in the social science domain.

The domain-specific track was established at the inception of CLEF in 2000 and was funded by the European Union from 2001-2004 (Kluck & Gey, 2001; Kluck, 2004). It is now continued at the GESIS German Social Science Information Centre (Bonn) in cooperation with the DELOS Network of Excellence on Digital Libraries.

The GIRT databases (now in version 4) are extracts from the German Social Science Information Centre's SOLIS (Social Science Literature) and SOFIS (Social Science Research Projects) databases from 1990-2000. In 2005, the Russian Social Science Corpus (RSSC) was added as a Russian-language test collection (94,581 documents), which was changed in 2006 to the INION ISISS corpus covering social sciences and economics in Russian. This year, another English-language social science collection was added. The second English collection is an extract from CSA's Sociological abstracts providing more documents and another thesaurus to the test bed.

In addition to the four test collections, various controlled vocabularies and mappings between vocabularies were made available. As is standard for the domain-specific track, 25 topics were prepared in German and then translated into English and Russian.

2 The Domain-Specific Task

The domain-specific track includes three subtasks:

- *Monolingual retrieval* against the German GIRT collection, the English GIRT and CSA Sociological Abstract collections, or the Russian INION ISISS collection;
- *Bilingual retrieval* from any of the source languages to any of the target languages;
- *Multilingual retrieval* from any source language to all collections / languages.

2.1 The Test Collections

In recent years, pseudo-parallel collections in German and English (GIRT) and one or two Russian test collections were provided (Kluck & Stempfhuber, 2005; Stempfhuber & Baerisch, 2006). This year, only one Russian but two English collections were provided.

Every test collection is in the format of a bibliographic database (records include title, author, abstract and source information) with the addition of subject metadata from controlled vocabularies.

German

The German GIRT collection (the social science German Indexing and Retrieval Testdatabase) is now used in its forth version (Girt-description, 2007) with 151,319 documents covering the years 1990-2000 using the German version of the Thesaurus for the Social Sciences. Almost all documents contain an abstract (145,941).

English

The English GIRT collection is a pseudo-parallel corpus to the German GIRT collection, providing translated versions of the German documents. It also contains 151,319 documents using the English version of the Thesaurus for the Social Sciences but only 17% (26,058) documents contain an abstract.

New additions this year were the documents from the social science database Sociological Abstracts from Cambridge Scientific Abstracts (CSA) with 20,000 documents, 94% of which contain an abstract. The documents were taken from the SA database covering the years 1994, 1995, and 1996. Additional to title and abstract, each document contains subject-describing keywords from the CSA Thesaurus of Sociological Indexing Terms and classification codes from the Sociological Abstracts classification.

Russian

For Russian retrieval, the INION corpus ISISS with bibliographic data from the social sciences and economics with 145,802 documents was once again used. ISISS documents contain authors, titles, abstracts (for 27% of the test collection or 39,404 documents) and keywords from the Inion Thesaurus.

2.2 Controlled Vocabularies

The GIRT collections have assigned descriptors from the GESIS IZ Thesaurus for the Social Sciences in German and English depending on the collection language. The CSA Sociological Abstracts documents contain descriptors from the CSA Thesaurus of Sociological Indexing Terms and the Russian ISISS documents are provided with Russian INION Thesaurus terms. GIRT documents also contain classification codes from the GESIS IZ classification and CSA SA documents from the Sociological Abstracts classification. Table 1 shows the distribution of subject-describing terms per document in each collection.

Collection	*GIRT-4 (German or English)*	*CSA Sociological Abstracts*	*INION ISISS*
Thesaurus descriptors / document	10	6.4	3.9
Classification codes / document	2	1.3	n/a

Table 1. Distribution of subject-describing terms per collection

Vocabulary mappings
Additional to the "mapping table" for the German and English terms from the GESIS IZ Thesaurus for the Social Sciences, which is really a translation, a bidirectional mapping between the GIRT and CSA Thesauri was provided.

Vocabulary mappings are one-directional, intellectually created term transformations between two controlled vocabularies. They can be used to switch from the subject metadata terms of one knowledge system to the other, enabling a retrieval system to treat the subject descriptions of two or more different collections as one and the same. This year's mappings were equivalence transformations, showing only term mappings that were found to be equivalent between two different controlled vocabularies.

We provided mappings between the German Thesaurus for the Social Sciences and the English CSA Thesaurus of Sociological Indexing Terms. Since the German Thesaurus for the Social Sciences exists in an English version as well, we also provided the mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms for monolingual retrieval.

An example for a mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms would be:

```
<mapping>
        <original-term>agricultural area </original-term>
        <mapped-term>Rural areas</mapped-term>
 </mapping>
```

This example shows that a mapping can overcome differences in technical language and the treatment of singular and plural in different controlled vocabularies.

2.3 Topic Preparation

As is standard for the CLEF domain-specific track, 25 topics were prepared.

For topic preparation we were supported by our colleagues from the GESIS Social Science Information Centre. As a special service to the social science community in Germany, the Information Centre biannually publishes updates on new entries in the SOLIS and SOFIS databases (from which the GIRT collections were generated). The specialized updates are prepared in 28 subject categories by subject specialists working at the Centre. Topics range from general sociology, family research, women's and gender studies, international relations, research on Eastern Europe to social psychology and environmental research. An overview of the service including the 28 topics can be found at the following URL:
http://www.gesis.org/en/information/soFid/index.htm.

We asked our colleagues to think of between 2-5 topics related to their subject area and potentially relevant in the years 1990-2000 (the coverage of our test collections). The suggestions from 15 different colleagues were then checked according to breadth, variance from previous years and coverage in the test collections. 25 topics were selected and edited into the CLEF topic XML format. Figure 1 is an example for a topic.

All topics were created in German and then consequently translated into English and Russian.

```
<top>
<num>192</num>
<EN-title>System change and family planning in East Germany</EN-title>
<EN-desc>Find documents describing birth trends and family planning since reunification in East Germany.</EN-desc>
<EN-narr>Of interest are documents on demographic changes which have taken place after 1989 in the territory of the former GDR as well as the slump in birth numbers, decline in marriages and divorces.</EN-narr>
</top>
```

Figure 1. Example topic in English

Table 2 lists all 25 topic titles in English to give a perspective on the variance in topics.

Sibling relations	Class-specific leisure behaviour
Unemployed youths without vocational training	Mortality rate
German-French relations after 1945	Economic elites in Eastern Europe and Russia
Multinational corporations	System change and family planning in East Germany
Partnership and desire for children	Gender and career chances
Torture in the constitutional state	Ecological standards in emerging or developing countries
Family policy and national economy	Integration policy
Women and income level	Tourism industry in Germany
Lifestyle and environmental behaviour	Promoting health in the workplace
Unstable employment situations	Economic situations of families
Value change in Eastern Europe	European climate policy
Migration pressure	Economic support in the East
Quality of life of elderly persons	

Table 2. Topic titles for domain-specific CLEF track 2007

To date, 200 topics have been created for the domain-specific track.

3 Overview of the 2007 Domain-Specific Track

More details of the individual runs and methods employed can be found in the corresponding articles by the participating groups.

3.1 Participants

Although 10 groups had registered for the domain-specific task, only 5 groups submitted runs. Four groups have submitted descriptions to the working notes so far (Clinchant and Renders, 2007; Fautsch et al., 2007; Kürsten and Eibl, 2007; Larson, 2007). Table 3 lists the participants.

Abbreviation	*Group Institution*	*Country*
Chemnitz	Media Informatics, Chemnitz University of Technology	Germany
Cheshire	School of Information, UC Berkeley	USA
Xerox	Xerox Research Centre Europe - Data Mining Group	France
Moscow	Moscow State University	Russia
Unine	Computer Science Department, University of Neuchatel	Switzerland

Table 3. Domain-specific track 2007 - participants

3.2 Submitted Runs

Experiments for all tasks (monolingual, bilingual and multilingual retrieval) were submitted to the track. Monolingual and bilingual experiments were equally attempted, whereas multilingual retrieval runs were only submitted by 2 groups. Russian remains slightly less popular than the other two languages. Table 4 provides the number of submitted runs per task, table 5 provides an overview over submitted runs per task per participant.

Task	Runs
Monolingual	
- against German	13
- against English	15
- against Russian	11
Bilingual	
- against German	14
- against English	15
- against Russian	9
Multilingual	9

Table 4. Submitted runs per task in the domain-specific track

Task	Participants (Runs)
Monolingual	
- against German	Chemnitz (3), Cheshire (2), Unine (4), Xerox (4)
- against English	Chemnitz (3), Cheshire (2), Moscow (2), Unine (4), Xerox (4)
- against Russian	Chemnitz (3), Cheshire (2), Moscow (2), Unine (4)
Bilingual	
- against German	Chemnitz (4), Cheshire (4), Xerox (6)
- against English	Chemnitz (3), Cheshire (4), Moscow (2), Xerox (6)
- against Russian	Chemnitz (3), Cheshire (4), Moscow (2)
Multilingual	Chemnitz (3), Cheshire (6)

Table 5. Submitted runs per task and participant

3.3 Relevance Assessments

In previous years, the domain-specific relevance assessments were administrated and overseen at least partly in-house at the Social Science Information Centre (using a self-developed Java-Swing program). This year all relevance assessments were administered and processed in the DIRECT system (Distributed Information Retrieval Evaluation Campaign Tool) provided by Giorgio M. Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova, Italy.

This provided tremendous assistance for the CLEF group at the Information Centre and was positively accepted by the five assessors. Some problems occurred because of bandwidth and execution problems, but overall the assessment stage went smoothly.

Documents were pooled using the top 100 ranked documents from each submission. Table 6 shows the pool sizes for each language.

German	16,288
English	17,867
Russian	14,473

Table 6. Pool sizes in the domain-specific track

For the German assessments, 652 documents per topic were judged on average and about 22% were found relevant. However, assessments vary from topic to topic. Figure 2 shows the German assessments per topic.

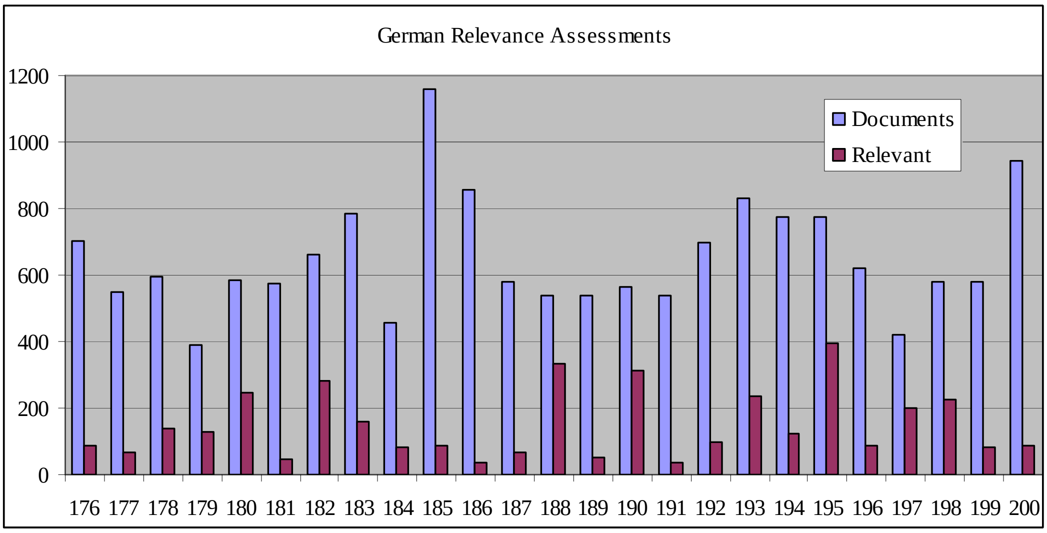

Figure 2. German assessments per topic

For the English assessments, 715 documents per topic were judged on average and about 25% were found relevant.

For the Russian assessments, 3 topics were found to have no relevant documents in the ISISS collection: 178, 181 and 191. For the assessments, 579 documents per topic were judged and only 10% were found relevant.

Figures 3 and 4 show the English and Russian relevance assessments numbers.

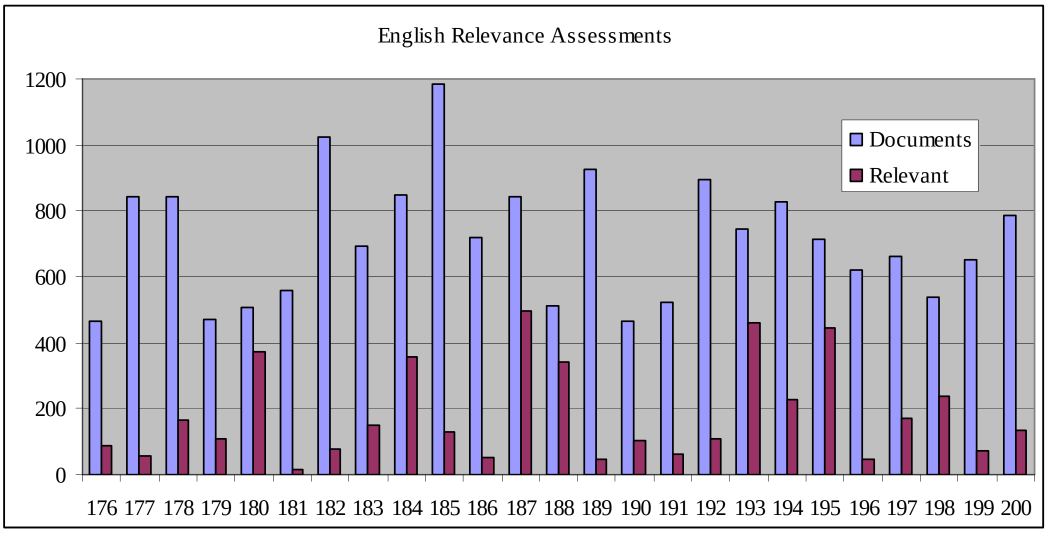

Figure 3. English assessments per topic

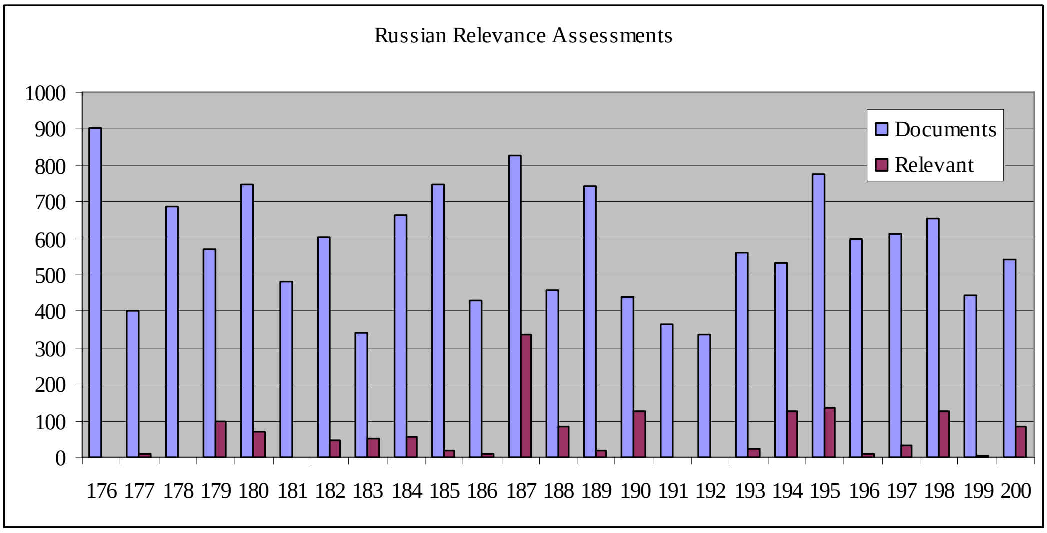

Figure 4. Russian assessments per topic

At first glance, several topics seem to yield particularly many relevant documents over all 3 languages despite different collections (e.g. 188, 190, 195) whereas others seem to yield particularly few (e.g. 181, 191). One explanation might be the timeliness and specificity of topics. The topics yielding many relevant topics (Quality of life of elderly persons, Mortality rate, Integration policy) seem to be rather broad and ongoing themes in the social science literature. The other two topics (Torture in the constitutional state, Economic elites in Eastern Europe and Russia) could be considered more specific and geared towards more recent time frames than others.

4 Domain-Specific Experiments

Every group used the controlled vocabularies and structured data in some facility or other. One point of emphasis was query expansion with the help of the subject description provided by the thesauri. However, the translation and mapping tables were also used as bilingual dictionaries for the cross-language experiments.

4.1 Retrieval models

The Chemnitz group (Kürsten and Eibl, 2007) used a redesigned version of their retrieval system based on the Lucene API and utilized two indices in retrieval: a structured index (taking the structure of the documents into account) and a plain index without considering the structure of the documents. To combine the two indices, a data fusion approach using the z-score introduced by the Unine group (Savoy, 2004) was employed. They found that the unstructured indexed outperformed the structured one.

The Berkeley group (Larson, 2007) used a probabilistic model employing a logistic regression algorithm successfully used for cross-language retrieval since TREC-2 and implemented it with the Cheshire retrieval system.

Unine (Fautsch et al., 2007) used several retrieval models for comparison purposes: the classical tf idf vector space model, probabilistic retrieval with the Okapi algorithm and four variants of the DFR (Divergence from Randomness) approach as well as a language modelling approach. Data fusion was applied using the z-score to combine these different models. They also compared word-based and n-gram indexing for retrieval with the Russian language corpus.

The Xerox group (Clinchant and Renders, 2007) used a language modelling approach for their retrieval experiments.

4.2 Language Processing for Documents and Queries

Standard language processing for documents and queries in the form of stopword-removal and stemming or normalization was employed by all groups. The Unine group successfully developed a new light-weight stemmer for the Russian language.

For the German language, Unine and Xerox used a decompounding module to split German compounds whereas Berkeley and Chemnitz did not.

4.3 Query Expansion

Three of the groups focused on query expansion in some way or another. Berkeley used a version of Entry Vocabulary Indexes (Gey et al., 2001) based on the same logistic regression algorithm as their retrieval system to associate title and description terms from topics with controlled vocabulary terms from documents. Another approach was a thesaurus-lookup where title and description words were looked up in a thesaurus that combined all subject-describing keywords from the different collections. The terms from the controlled vocabularies were added to the query. As part of its standard retrieval process, the Cheshire system also implemented a blind feedback algorithm based on the Robertson and Sparck Jones term weights. Whereas the Entry Vocabulary Index approach worked better for the English target language, the thesaurus look-up worked better for German and Russian.

Unine used the Thesaurus for the Social Sciences to enhance queries with terms from the thesaurus. Thesaurus entries were indexed as documents and retrieved in response to query terms, then simply added to the query. They also used blind query feedback with Rocchio’s formula as well as an idf-based approach described in Abdou and Savoy (2007). The blind feedback approach improved the average precision of results, whereas the thesaurus expansion did not.

Xerox used lexical entailment to provide query expansion whereby a language modelling approach is employed to find similar terms from corpus documents in relation to query terms. They found that this approach outperformed simple blind feedback but a combined approach worked best.

4.4 Translation

Another focus of research was query translation, where the provided mapping tables were utilized as bilingual dictionaries.

Berkeley used the commercially available LEC Power Translator program for translation in all languages.

Chemnitz implemented a translation-plug-in to their Lucene retrieval system utilizing well-known freely-available translation services like Babel Fish, Google Translate, PROMT and Reverso. They also used the bilingual mapping table from the thesauri for translation.

Finally, Xerox compared their Statistical Machine Translation System MATRAX with a sophisticated language-model-based approach of dictionary adaptation. Dictionary adaptation attempts to select one out of several translation possibilities for a term using a bilingual dictionary and calculating the probability of a target term given the language context of the source query term. They found that this approach worked well compared to the statistical machine translation system tested.

5 Results

In the Appendix of this volume, mean average precision numbers (MAP) for each run per task and recall-precision graphs for the top-performing runs for each task are listed.

6 Outlook

This year's experiments have shown that leveraging a controlled vocabulary for query expansion or translation can improve results in structured test collections. A new collection and new vocabulary (CSA Sociological Abstracts) was added and a mapping table between the CSA Thesaurus and the GIRT Thesaurus provided for experiments. As new collections are added and distributed search across several collections becomes more common, the seamless switching between controlled vocabularies becomes crucial to utilize expansion and translation techniques developed for individual collections.

For this purpose, several resources for terminology mapping have been developed at the German Social Science Information Centre (KoMoHe Project Website, 2007). Among them are over 40 bidirectional mappings between various controlled vocabularies. A web service to retrieve mapped terms is being developed. Besides the expansion of test collections, these vocabulary mapping services could be a future branch of research for the domain-specific track within CLEF.

Acknowledgements

We would like to thank Cambridge Scientific Abstracts for providing the documents for the new Sociological Abstracts test collection.

We greatly acknowledge the support of Natalia Loukachevitch and her colleagues from the Research Computing Center of M.V. Lomonosov Moscow State University in translating the topics into Russian as well as providing parts of the Russian relevance assessments.

Very special thanks also to Giorgio Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova for providing the DIRECT system and all their help in the assessments process and for providing the graphs and numbers for the results analysis.

Claudia Henning and Jeof Spiro did the German and English assessments; Monika Gonser and Oksana Schäfer provided the rest of the Russian assessments.

References

Abdou, S. and Savoy J. (2007). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. Information Processing & Management, to appear.

Stephane Clinchant and Jean-Michel Renders (2007). XRCE's Participation to CLEF 2007 Domain-specific Track. This volume.

Claire Fautsch, Ljiljana Dolamic, Samir Abdou and Jacques Savoy (2007). Domain-Specific IR for German, English and Russian Languages. This volume.

Fredric Gey, Michael Buckland, Aitao Chen, and Ray Larson (2001). Entry vocabulary – a technology to enhance digital search. In Proceedings of HLT2001, First International Conference on Human Language Technology, San Diego, pages 91–95, March 2001.

Girt Description (2007). GIRT - Mono- and Cross-language Domain-Specific Information Retrieval (GIRT4). http://www.gesis.org/en/research/information_technology/girt4.htm

Michael Kluck and Frederik C. Gey (2001). The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval . In: Carol Peters (ed.): Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Information Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. Berlin/Heidelberg/New York: Springer 48-56 (Lecture Notes in Computer Science, 2069)

Michael Kluck (2004). The GIRT Data in the Evaluation of CLIR Systems – from 1997 until 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (Eds..) Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Berlin/Heidelberg/New York: Springer 2004, 379-393 (Lecture Notes in Computer Science, 3237)

Michael Kluck and Maximilian Stempfhuber (2005). Domain-Specific Track CLEF 2005: Overview of Results and Approaches, Remarks on the Assessment Analysis. Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/kluck05.pdf

KoMoHe Project Website (2007). Competence Center Modeling and Treatment of Semantic Heterogeneity. http://www.gesis.org/en/research/information_technology/komohe.htm

Jens Kürsten and Maximilian Eibl (2007). Domain-Specific Cross Language Retrieval: Comparing and Merging Structured and Unstructured Indices. This volume.

Ray Larson (2007). Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language Retrieval. This volume.

Savoy, Jacques (2004). Data Fusion for Effective European Monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/22.pdf

Maximilian Stempfhuber and Stefan Baerisch (2006). Domain-Specific Track CLEF 2005: Overview of Results and Approaches, Remarks on the Assessment Analysis. Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/stempfhuberOCLEF2006.pdf