What happened in CLEF 2005?

Introduction to the Working Notes

Carol Peters
Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy
carol.peters@isti.cnr.it

Each year, the Cross-Language Evaluation Forum (CLEF) organises a series of evaluation tracks designed to test different aspects of mono- and cross-language information retrieval system development. From the very beginning the intention has been to encourage experimentation with all kinds of multilingual information access – from the development of systems for monolingual retrieval operating on many languages to the implementation of complete multilingual multimedia search services. In addition, CLEF aims at encouraging contacts between the R&D and the industrial communities and promoting the take-up and porting of research results into real world applications.

These Working Notes contain descriptions of the experiments conducted within CLEF 2005 – the sixth in a series of annual system evaluation campaigns[1]. The results of the experiments will be presented and discussed in the CLEF 2005 Workshop, 21-23 September, Vienna, Austria. The final papers - revised and extended as a result of the discussions at the Workshop - together with a comparative analysis of the results will appear in the CLEF 2005 Proceedings, to be published by Springer in their Lecture Notes for Computer Science series.

Up until this year the Working Notes were prepared in printed form, in a limited number, and distributed at the Workshop to all the participants. They were also posted on the CLEF web-site, immediately following the workshop, to facilitate dissemination to the interested research community. However, as participation in CLEF has increased over the years, the size of the Working Notes has grown accordingly. Last year, we printed two volumes for a total of almost 1000 pages. This year we decided on a new scheme: the Working Notes containing full reports of all experiments would be published in electronic format only. The CLEF 2005 Working Notes have thus been posted it on the CLEF website and have also been inserted in the DELOS Digital Library, accessible at http://delos-dl.isti.cnr.it. A limited number have also been prepared on CD for distribution to workshop participants together with a set of extended abstracts containing brief descriptions of all the experiments.

Although the form of the Working Notes has changed, the content remains the same. They are divided into eight sections, corresponding to the CLEF 2005 evaluation tracks. In addition appendices are included containing run statistics for the Ad Hoc, Domain-Specific, GeoCLEF and CL-SR tracks, and a list of all participating groups showing in which track they took part.

The main features of the 2005 campaign are briefly outlined here below in order to provide the necessary background to the experiments reported in the rest of the Working Notes.

1. Tracks and Tasks in CLEF 2005

Over the years CLEF has gradually increased the number of different tracks and tasks offered in order to facilitate experimentation with all kinds of multilingual information access. CLEF 2005 offered eight tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual textual document retrieval on news collections (Ad Hoc)
- mono- and cross-language information on structured scientific data (Domain-Specific)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SR)
- multilingual retrieval of Web documents (WebCLEF)
- cross-language geographical retrieval (GeoCLEF)

[1] CLEF is included in the activities of the DELOS Network of Excellence on Digital Libraries, funded by the Sixth Framework Programme of the European Commission. For information on DELOS, see www.delos.info.

Cross-Language Text retrieval (Ad Hoc): As in past years, the CLEF 2005 ad hoc track was structured in three tasks, testing systems for monolingual (querying and finding documents in one language), bilingual (querying in one language and finding documents in another language) and multilingual (querying in one language and finding documents in multiple languages) retrieval. The monolingual and bilingual tasks were principally offered for Bulgarian, French, Hungarian and Portuguese target collections. Additionally, in the bilingual task only, newcomers (i.e. groups that had not previously participated in a CLEF cross-language task) or groups using a "new-to-CLEF" query language could choose to search the English document collection. The Multilingual task was based on the CLEF 2003 multilingual-8 test collection which contained news documents in eight languages: Dutch, English, French, German, Italian, Russian, Spanish, and Swedish. There were two subtasks. a traditional multilingual retrieval task requiring participants to carry out retrieval and merging (Multi-8 Two-Years-On), and a new task focussing only on the multilingual results merging problem using standard sets of ranked retrieval output (Multi-8 Merging Only).

Cross-Language Scientific Data Retrieval (Domain-Specific): This track studied retrieval in a domain-specific context using the GIRT-4 German/English social science database and the Russian Social Science Corpus (RSSC). Multilingual controlled vocabularies (German-English, English-German, German-Russian, English-Russian) were available. Monolingual and cross-language tasks were offered. Topics were prepared in English, German and Russian. Participants could make use of the indexing terms inside the documents and/or the Social Science Thesaurus provided, not only as translation means, but also for tuning relevance decisions of their system.

Interactive CLIR (iCLEF): The challenge in this track is to build a system that will allow real people to find information that is written in languages that they have not mastered, and then measure how well representative users are able to use the system that has been built. This year, iCLEF focused on problems of cross-language question answering and image retrieval from a user-inclusive perspective. Participating groups were to adapt a shared user study design to test a hypothesis of their choice, comparing reference and contrastive systems.

Multilingual Question Answering (QA@CLEF): Monolingual (non-English) and cross-language QA systems were tested. Combinations between nine target collections (Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese and Spanish) and questions in ten languages (the target languages plus Indonesian) were explored. Both factoid and definition questions were provided as input; a subset of the factoid questions were temporally restricted.

Cross-Language Retrieval in Image Collections (ImageCLEF): The aim of this track was to explore the use of both text and content-based retrieval methods for cross-language image retrieval. Three main tasks were offered: ad-hoc retrieval from a historic photographic collection, ad-hoc retrieval from a medical collection, and an automatic image annotation task.

Cross-Language Speech Retrieval (CL-SR): The focus this year was on searching spontaneous speech from oral history interviews rather than news broadcasts. The test collection created for the track is a subset of a large archive of videotaped oral histories from survivors, liberators, rescuers and witnesses of the Holocaust created by the Survivors of the Shoah Visual History Foundation (VHF). Automatic Speech Recognition (ASR) transcripts and both automatically assigned and manually assigned thesaurus terms were available as part of the collection. Topics were translated from English into Czech, French, German and Spanish to facilitate cross-language experimentation.

The final two tracks were introduced for the first time in CLEF 2005 as experimental pilot tracks.

Multilingual Web Retrieval (WebCLEF): The aim of this track was to evaluate systems that address multilingual information needs on the web. Three tasks were organized: mixed monolingual, multilingual, and bilingual English to Spanish, with 242 homepage and 305 named page finding queries for the first two tasks, and 67 homepage and 67 named page finding tasks for the third task.

Cross-Language Geographical Retrieval (GeoCLEF): The aim of GeoCLEF was to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants were offered a TREC-style ad hoc retrieval task based on existing CLEF collections.

Details on the technical infrastructure and the organisation of these tracks can be found in the track overview reports in this volume, collocated at the beginning of the relevant sections.

2. Document Collections

Seven different document collections have been used in CLEF 2005 to build the test collections:

- CLEF multilingual comparable corpus of more than 2 million news docs in 12 languages (see Table 1)
- The GIRT-4 social science database in English and German and the Russian Social Science Corpus
- St Andrews historical photographic archive
- CasImage radiological medical database with case notes in French and English
- IRMA collection in English and German for automatic medical image annotation
- Malach collection of spontaneous conversational speech derived from the Shoah archives
- EuroGOV, a multilingual collection of about 2M webpages crawled from European governmental sites.

Table 1: Sources and dimensions of the CLEF 2005 multilingual comparable corpus

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens)[2]	Median Size of Docs (Features)
Bulgarian: Sega 2002	2005	120	33,356	NA	NA	NA
Bulgarian: Standart 2002	2005	93	35,839	NA	NA	NA
Dutch: Algemeen Dagblad 94/95	2001	241	106483	1282	166	112
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153	354	203
English: LA Times 94	2000	425	113005	2204	421	246
English: Glasgow Herald 95	2003	154	56472	2219	343	202
Finnish: Aamulehti late 94/95	2002	137	55344	1712	217	150
French: Le Monde 94	2000	158	44013	1994	361	213
French: ATS 94	2001	86	43178	1683	227	137
French: ATS 95	2003	88	42615	1715	234	140
German: Frankfurter Rundschau94	2000	320	139715	1598	225	161
German: Der Spiegel 94/95	2000	63	13979	1324	213	160
German: SDA 94	2001	144	71677	1672	186	131
German: SDA 95	2003	144	69438	1693	188	132
Hungarian: Magyar Hirlap 2002	2005	105	49,530	NA	NA	NA
Italian: La Stampa 94	2000	193	58051	1915	435	268
Italian: AGZ 94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Portuguese: Público 1994	2004	164	51751	NA	NA	NA
Portuguese: Público 1995	2004	176	55070	NA	NA	NA
Portuguese: Folha 94	2005	108	51,875	NA	NA	NA
Portuguese: Folha 95	2005	116	52,038	NA	NA	NA
Russian: Izvestia 95	2003	68	16761	NA	NA	NA
Spanish: EFE 94	2001	511	215738	2172	290	171
Spanish: EFE 95	2003	577	238307	2221	299	175
Swedish: TT 94/95	2002	352	142819	2171	183	121

SDA/ATS/AGZ = Schweizerische Depeschenagentur (Swiss News Agency)
EFE = Agencia EFE S.A (Spanish News Agency)
TT = Tidningarnas Telegrambyrå (Swedish newspaper)

[2] The number of tokens extracted from each document can vary slightly across systems, depending on the respective definition of what constitutes a token. Consequently, the number of tokens and features given in this table are approximations and may differ from actual implemented systems.

Two new collections – Bulgarian and Hungarian newspapers for 2002 - were added to the multilingual corpus this year. Moreover, the Portuguese collection was expanded with the addition of a Brazilian newspaper: Folha. The multilingual corpus thus now contains approximately 2 million news documents in twelve languages, for 1994-1995**:** Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish and Swedish, and for 2002: Bulgarian and Hungarian. Table 1 gives the main specifics. Parts of this collection were used by the Ad Hoc (all languages except Russian), Question Answering (all languages except Hungarian, Russian and Swedish), Interactive (English and French) and GeoCLEF (English and German) tracks in CLEF 2005.

The domain-specific track used two collections: the GIRT-4 collection derived from the GIRT (German Indexing and Retrieval Test) social science database and RSSC (the Russian Social Science Corpus) GIRT-4 consists of over 150,000 documents includes a pseudo-parallel English/German corpus. Controlled vocabularies in German-English and German-Russian were also made available to the participants in this track. RSSC contains approximately 95,000 Russian social science documents.

The ImageCLEF track used three distinct collections: a collection of approximately 28,000 historic photographs with associated textual captions and metadata provided by St Andrews University, Scotland; a collection of about 9,000 medical images with French/English case notes made available by the University Hospitals, Geneva., and the IRMA database of 10,000 medical images made available by the IRMA group, Aachen University of Technology (RWTH).

The speech retrieval track used the MALACH collection extracted from the Shoah archives. The sub-collection used in CLEF 2005 contained 8,104 manually identified segments from 272 English interviews (589 hours).

The WebCLEF track used a collection crawled from European governmental sites, called EuroGOV. This collection consists of more than 3.35 million pages from 27 primary domains. The most frequent languages are Finnish (20%), German (18%), Hungarian (13%), English (10%), and Latvian (9%).

3. Participation

A total of 74 groups submitted runs in CLEF 2005, as opposed to the 54 groups of CLEF 2004: 43(37) from Europe, 19(12) from N.America; 10(5) from Asia and 1 each from S.America and Australia. Last years' figures are given between brackets. The breakdown of participation of groups per track is as follows: Ad Hoc 23; Domain-Specific 8; iCLEF 5; QAatCLEF 24; ImageCLEF 24; CL-SR 7; WebCLEF 11; GeoCLEF 12. As in previous years, participating groups consist of a nice mix of new-comers (26) and groups that had participated in one or more previous editions (48). A list of groups and indications of the tracks in which they participated in is given in Appendix to these Working Notes.

The introduction of new tracks this year has clearly had a big impact both with respect to numbers and also regarding expertise – making CLEF an increasingly multidisciplinary forum. Figure 1 shows the growth in participation over the years and Figure 2 shows the shift in focus as new tracks have been added.

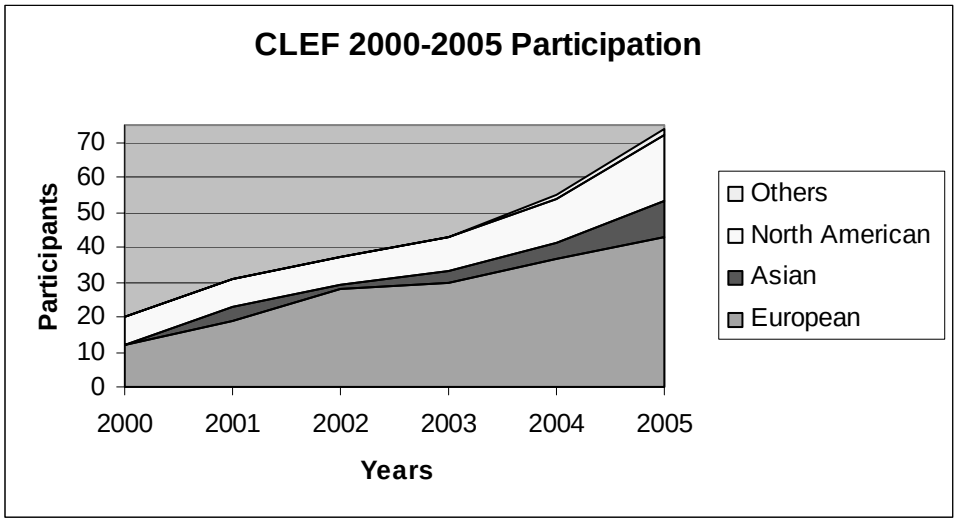

Figure 1. CLEF 2000 – 2005: Increase in Participation

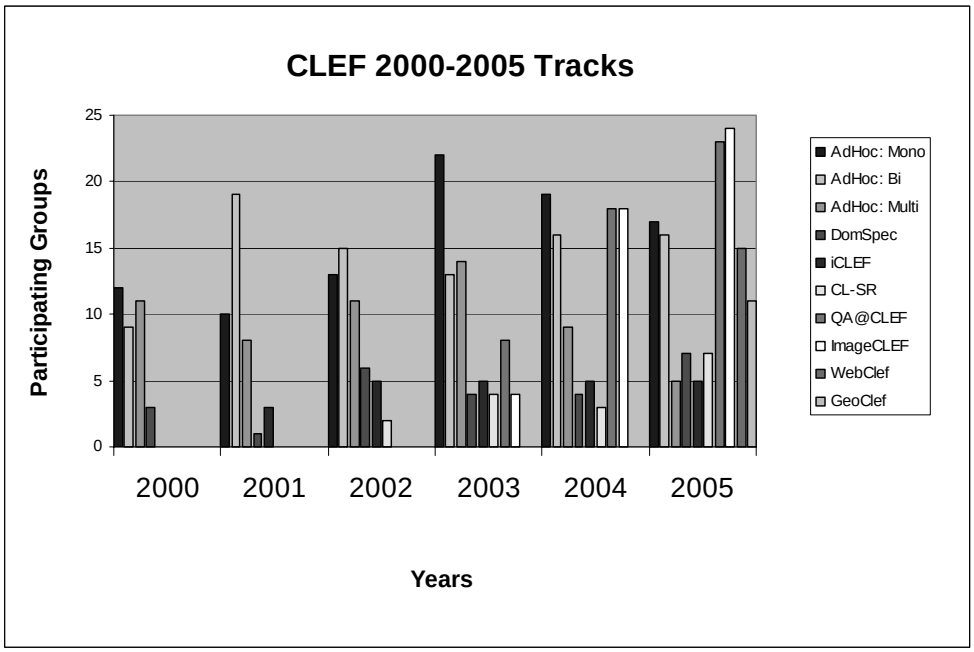

Figure 2. CLEF 2000 – 2005: Increase in Tracks

4. Workshop

CLEF aims at creating a strong CLIR/MLIR research and development community. The Workshop plays an important role by providing the opportunity for all the groups that have participated in the evaluation campaign to get together comparing approaches and exchanging ideas. The work of the groups participating in this year's campaign will be presented in plenary paper and poster sessions. There will also be break-out sessions for more in-depth discussion of the results of individual tracks and intentions for the future. The final sessions will include discussions on ideas for new tracks in future campaigns. Overall, the Workshop should provide an ample panorama of the current state-of-the-art and the latest research directions in the multilingual information retrieval area. I very much hope that it will prove an interesting, worthwhile and enjoyable experience to all those who participate.

The final programme and the presentations at the Workshop will be posted on the CLEF website at http://www.clef-campaign.org.

Acknowledgements

It would be impossible to run the CLEF evaluation initiative and organize the annual workshops without considerable assistance from many groups.. CLEF is organized on a distributed basis, with different research groups being responsible for the running of the various tracks. My gratitude goes to all those who have been involved in the coordination of the 2005 campaigns. A list of the main institutions involved is given on the following page. Here below, let me thank those responsible for the coordination of the different tracks:

- Giorgio Di Nunzio, Nicola Ferro and Gareth Jones for the Ad Hoc Track
- Michael Kluck and Natalia Loukachevitch for the Domain-Specific track
- Julio Gonzalo, Paul Clough and Alessandro Vallin for iCLEF
- Bernardo Magnini, Alessandro Vallin, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos and Richard Sutcliffe for QA@CLEF
- Paul Clough, Henning Müller, Thomas Deselaers , Michael Grubinger, Thomas Lehmann, Jeffery Jensen, and William Hersh for ImageCLEF
- Ryen W. White, Douglas W. Oard, Gareth J. F. Jones, Dagobert Soergel, Xiaoli Huang for CL-SR
- Börkur Sigurbjörnsson, Jaap Kamps, Maarten de Rijke for Web-CLEF
- Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho and Paul Clough for GeoCLEF

In addition, I must express my appreciation to Diana Santos and her colleagues at Linguateca in Norway and Portugal, for all their efforts aimed at supporting the inclusion of Portuguese in CLEF activities. These Working Notes include a paper by Diana and Nuno Cardoso which reflects on the challenges that have to be addressed when including a language into CLEF evaluation activities. I also thank all those colleagues who have helped us

by preparing topic sets in different languages and in particular the NLP Lab. Dept. of Computer Science and Information Engineering of the National Taiwan University for their work on Chinese..

I should also like to thank the members of the CLEF Steering Committee who have assisted me with their advice and suggestions throughout this campaign.

Furthermore, I gratefully acknowledge the support of all the data providers and copyright holders, and in particular:

- The Los Angeles Times, for the American-English data collection
- SMG Newspapers (The Herald) for the British-English data collection
- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French data
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections
- InformationsZentrum Sozialwissen-schaften, Bonn, for the GIRT database
- SocioNet system for the Russian Social Science Corpora
- Hypersystems Srl, Torino and La Stampa, for the Italian data
- Agencia EFE S.A. for the Spanish data
- NRC Handelsblad, Algemeen Dagblad and PCM Landelijke dagbladen/Het Parool for the Dutch newspaper data
- Aamulehti Oyj and Sanoma Osakeyhtiö for the Finnish newspaper data
- Russika-Izvestia for the Russian newspaper data
- Público, Portugal, and Linguateca for the Portuguese (PT) newspaper collection
- Folha, Brazil, and Linguateca for the Portuguese (BR) newspaper collection
- Tidningarnas Telegrambyrå (TT) SE-105 12 Stockholm, Sweden for the Swedish newspaper data
- Schweizerische Depeschenagentur, Switzerland, for the French, German and Italian Swiss news agency data
- Ringier Kiadoi Rt. [Ringier Publishing Inc.].and the Research Institute for Linguistics, Hungarian Acad. Sci. for the Hungarian newspaper documents
- Sega AD, Sofia; Standart Nyuz AD, Sofia, and the BulTreeBank Project, Linguistic Modelling Laboratory, IPP, Bulgarian Acad. Sci, for the Bulgarian newspaper documents
- St Andrews University Library for the historic photographic archive
- University and University Hospitals, Geneva, Switzerland and Oregon Health and Science University for the ImageCLEFmed Radiological Medical Database
- Aachen University of Technology (RWTH), Germany for the IRMA database of annotated medical images
- The Survivors of the Shoah Visual History Foundation, and IBM for the Malach spoken document collection

Without their contribution, this evaluation activity would be impossible.

Last and not least, I should like to express our gratitude to both Francesca Borri and Valeria Quochi in Pisa and Andreas Rauber and Rudolf Mayer, Technical University Vienna, for their assistance in the organisation of the CLEF 2005 Workshop.

Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa. The following institutions have contributed to the organisation of the different tracks of the CLEF 2005 campaign:

- Centre for the Evaluation of Human Language and Multimodal Communication Technologies (CELCT), Trento, Italy
- Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Trento, Italy
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, USA
- Department of Computer Science, University of Helsinki
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- Department of Information Engineering, University of Padua, Italy
- Department of Information Studies, University of Sheffield, UK
- Evaluations and Language Resources Distribution Agency Sarl, Paris, France
- German Research Centre for Artificial Intelligence, DFKI, Saarbrücken,
- Information and Language Processing Systems, University of Amsterdam, Netherlands
- InformationsZentrum Sozialwissenschaften, Bonn, Germany
- Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguateca, Sintef, Oslo, Norway; University of Minho, Braga, Portugal
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
- National Institute of Standards and Technology, Gaithersburg MD, USA
- Oregon Health and Science University, USA
- Research Computing Center of Moscow State University
- Research Institute for Linguistics, Hungarian Academy of Sciences
- School of Computing, Dublin City University, Ireland
- UC Data Archive and School of Information Management and Systems, UC Berkeley, USA
- University Hospitals and University of Geneva, Switzerland

CLEF Steering Committee

Maristella Agosti, University of Padova, Italy
Eija Airio, University of Tampere, Finland
Martin Braschler, Zurich, Switzerland
Amedeo Cappelli, ISTI-CNR & CELCT, Italy
Hsin-Hsi Chen, National Taiwan University, Taipei, Taiwan
Khalid Choukri, Evaluations and Language resources Distribution Agency, Paris, France
Paul Clough, University of Sheffield, UK
David A. Evans, Clairvoyance Corporation, USA
Marcello Federico, ITC-irst, Trento, Italy
Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
Norbert Fuhr, University of Duisburg, Germany
Frederic C. Gey, U.C. Berkeley, USA
Julio Gonzalo, LSI-UNED, Madrid, Spain
Donna Harman, National Institute of Standards and Technology, USA
Gareth Jones, Dublin City University, Ireland
Franciska de Jong, University of Twente, Netherlands
Noriko Kando, National Institute of Informatics, Tokyo, Japan
Jussi Karlgren, Swedish Institute of Computer Science, Sweden
Michael Kluck, German Institute for International and Security Affairs, Berlin, Germany
Natalia Loukachevitch, Moscow State University, Russia
Bernardo Magnini, ITC-irst, Trento, Italy
Paul McNamee, Johns Hopkins University, USA
Henning Müller, University & University Hospitals of Geneva, Switzerland
Douglas W. Oard, University of Maryland, USA
Maarten de Rijke, University of Amsterdam, Netherlands
Jacques Savoy, University of Neuchatel, Switzerland
Peter Schäuble, Eurospider Information Technologies, Switzerland
Richard Sutcliffe, University of Limerick, Ireland
Max Stempfhuber, Informationszentrum Sozialwissenschaften Bonn, Germany
Hans Uszkoreit, German Research Center for Artificial Intelligence (DFKI), Germany
Felisa Verdejo, LSI-UNED, Madrid, Spain
José Luis Vicedo, University of Alicante, Spain
Ellen Voorhees, National Institute of Standards and Technology, USA
Christa Womser-Hacker, University of Hildesheim, Germany