What happened in CLEF 2004?
Introduction to the Working Notes

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

These Working Notes contain descriptions of the experiments conducted within CLEF 2004 – the fifth in a series of annual system evaluation campaigns organised by the Cross-Language Evaluation Forum[1]. The results of the experiments will be presented and discussed in the CLEF 2004 Workshop, 15-17 September, Bath, UK. 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 2004 Proceedings, to be published by Springer in their Lecture Notes for Computer Science series.

CLEF organises a series of evaluation tracks designed to test different aspects of mono- and cross-language information retrieval system development with the main focus on European languages. The objective is to provide an infrastructure that facilitates experimentation with all kinds of multilingual information access – from the development of procedures 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 application communities and promoting the industrial take-up of research results.

The main features of the 2004 campaign are briefly outlined here below in order to provide the necessary background to the experiments reported in this volume.

1. Tracks and Tasks in CLEF 2004

In recent years, CLEF has distinguished between core tracks, which were those offered regularly in each campaign (the monolingual, bilingual, multilingual and domain-specific tracks), and additional tracks, which were organised on an experimental basis with the objective of identifying new requirements and appropriate methodologies for their testing in a cross-language context. This distinction no longer held in 2004. The great success of the so-called additional tracks in CLEF 2003, and in particular of the tracks that tested systems for question answering and image retrieval, has led to their inclusion as regular tracks this year. This has meant that CLEF 2004 marks a breaking point with respect to the previous campaigns. The focus is no longer on multilingual document retrieval but has diversified to include different kinds of text retrieval across languages (from documents to exact answers) and retrieval on different kinds of media (not just text but image and speech as well).

CLEF 2004 thus offered six tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad-hoc)
- mono- and cross-language domain-specific retrieval (GIRT)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval on image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SDR)

In the following sub-sections, I will describe the organisation of the first two tracks. Details on the other tracks can be found in the track overviews reported in this volume and collocated at the beginning of the relevant sections.

1.1 Multilingual/Bilingual/Monolingual Document Retrieval: One of the principal objectives when CLEF began was to encourage developers to build truly multilingual retrieval systems capable of using a query in one language to find relevant documents in any of the languages contained in a collection, listing the results in a single, ranked list. For several years, the multilingual track has been considered to be the main track in CLEF, and the bilingual and monolingual tracks as steps leading system developers towards this goal. These tracks are known collectively as the CLEF ad hoc tracks.

[1] CLEF 2004 is included in the activities of the DELOS Network of Excellence on Digital Libraries, funded by the Sixth Framework Programme of the European Commission. DELOS is an "old" friend of CLEF, having promoted the first two campaigns in 2000 and 2001. For information on DELOS, see www.delos.info.

The multilingual task was thus made progressively more difficult in each campaign reaching a climax in CLEF 2003 where two distinct tasks were offered: multilingual-4 and multilingual-8. The aim was to facilitate first-time participation in this track with multilingual-4, but also to offer a stimulating task for more experienced groups with multilingual-8. The collection for multilingual-4 contained English, French, German and Spanish documents. This was increased to include Dutch, Finnish, Italian and Swedish documents for multilingual-8. The two tracks were a considerable success – with 14 participants overall. The bilingual tasks in 2003 were also particularly challenging. The main objective was to encourage the tuning of systems running on "unusual" language pairs. For this reason, experiments were solicited for specific source -> target languages pairs: Italian -> Spanish, German -> Italian, French -> Dutch, Finnish -> German. At the very last moment, Russian was also included as a possible target language. The CLEF 2003 monolingual track tested system performance on eight European languages (English was excluded).

With CLEF 2003 we felt that we had achieved an important goal. We had shown that fully multilingual retrieval could be (almost) as effective as bilingual (L1 -> L2) retrieval and that systems are able to adapt and reengineer rapidly and effectively to process new languages as the need arises. We had also created an important test collection for system benchmarking purposes[2]. We thus decided to reduce the ad hoc tracks in CLEF 2004 to leave more space for other types of multilingual/cross-language information retrieval experiments[3].

For this reason, the document collection used in the CLEF2004 ad hoc tracks contained just English, Finnish, French, Russian and Portuguese texts, with Portuguese a new acquisition.

The multilingual task solicited experiments retrieving documents from a collection containing documents in four of these languages (Portuguese excluded).

The bilingual track again imposed particular conditions on the source -> target language pairs accepted:

- Italian/French/Spanish/Russian queries -> Finnish target collection
- German/Dutch/Finnish/Swedish queries -> French target collection
- Any query language -> Russian target collection
- Any query language -> Portuguese target collection

As always, newcomers to a CLEF cross-language task or groups using a very new topic language were allowed to submit runs to the English target collection.

The monolingual track offered testing for four languages: Finnish, French, Russian and Portuguese.

For each of the above tracks, the participating systems constructed their queries (automatically or manually) from a common set of topics, created to simulate user information needs. Each topic consisted of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. Topic sets were produced by native speakers in the five document languages and additionally for Amharic, Bulgarian, Chinese, Dutch, German, Italian, Japanese, Spanish and Swedish. All topic languages were used by at least one group. As in previous years, a condition was that, for each task attempted, a mandatory run using the title and description fields had to be submitted. The objective is to facilitate comparison between the results of different systems.

Relevance assessment was also performed in all cases by native speakers. The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall techniques are calculated using pooling techniques. The results submitted by the participating groups were used to form a pool of documents for each topic and for each language by collecting the highly ranked documents from all submissions. The results were then analysed and run statistics produced and distributed. These are given in Appendix A.

1.2 Domain-Specific Information Retrieval: As always, the aim in 2004 for this track has been to study retrieval on other types of collections, serving different kinds of information needs. The information provided by domain-specific scientific documents is highly targeted and contains much terminology. The domain-specific track offered mono- and bilingual tasks on the GIRT4 collection of social science documents. 25 topics were prepared in three languages: English, German and Russian. The topics were created and relevance assessments were performed by domain experts.

[2] In the near future, this first CLEF multilingual test-suite will be made publicly available on the ELDA catalog (see www.elda.fr)

[3] This decision was also motivated by financial considerations. The contract with the European Commission that funded much of CLEF 2002 and 2003 was concluded in March 2004. Since then, CLEF has received only a small amount of funds from the DELOS Network; most of the work for CLEF 2004 has been conducted by groups working on a non-funded basis.

Details on the technical infrastructure and the organisation of the other four tracks (iCLEF, QA@CLEF, ImageCLEF, CL-SDR) can be found in the track overview reports in this volume, collocated at the beginning of the relevant sections.

2. Document Collections

As already mentioned, a new collection was added to the main CLEF multilingual comparable corpus this year: Público a Portuguese daily newspaper[4]. The multilingual corpus thus now contains nearly 1.8 million news documents from the same time period (1994-1995) in ten languages**:** Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish and Swedish. Table 1 gives the main specifics of this collection.

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens)[5]	Median Size of Docs (Features)
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
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
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)
NA = Not Available at this moment

Table 1: Sources and dimensions of the main CLEF 2004 multilingual document collection

The domain-specific track used the same collection as in CLEF 2003, the GIRT-4 collection derived from the GIRT (German Indexing and Retrieval Test) social science database. This corpus 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.

[4] The final section of this volume contains a paper by the group that introduced Portuguese into CLEF describing the efforts necessary to add a new language to the CLEF collection.

[5] 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.

The ImageCLEF track used two distinct collections: a collection of historic photographs provided by St Andrews University, Scotland, and a collection of medical images with French and English case notes made available by the University Hospitals, Geneva.

The cross-language spoken document retrieval track (CL-SDR) used speech transcriptions in English from the TREC-8 and TREC-9 SDR tracks, supplied by the National Institute of Standards and Technology (NIST), USA.

Table 2 shows which collections were used by the tracks in CLEF 2004.

TRACK/TASK	**DE**	**Brit-EN**	**US-EN**	**ES**	**FI**	**FR**	**IT**	**NL**	**PT**	**RU**	**SV**	**GIRT-4**	**Hist. Phot**	**Med. Image**	**TREC-SDR**
multilingual: 95 data only		X			X	X				X					
bilingual: 95 data only – according to task		only new-comer			X	X			X	X					
monolingual: 95 data only – according to task					X	X			X	X					
GIRT – mono- and bilingual												X			
iCLEF: 94 & 95 data according to task		X	X	X		X									
QA@CLEF: 94 & 95 data according to task	X	X	X	X		X	X	X	X						
ImageCLEF: 95 data according to task													X	X	
CL-SDR: English transcriptions															X

Table 2: Data collections used in CLEF 2004

3. Participation

A total of 55 groups submitted results in CLEF 2004: 36 from Europe, 13 from N.America; 4 from Asia and one mixed European/Asian group. This is a considerable increase on the 42 groups of CLEF 2003. 11 groups consisted of a collaboration between researchers from different institutions. A disappointment was that only six groups this year included representatives from industry. Many groups participated in more than one track. The breakdown of participation of groups per track is as follows: multilingual: 9; bilingual: 16; monolingual: 19; GIRT: 4; iCLEF: 5; QAatCLEF: 18; ImageCLEF: 18. Unfortunately, the CL-SDR track has problems this year with few participants and the results were considered to be of no great significance. As in previous years, participating groups consist of a nice mix of new-comers (23) and groups that had participated in one or more previous editions (32). Table 3 lists the participating groups – the asterisks indicate the number of times a group has participated in previous editions of CLEF. The six groups with four asterisks have taken part in all editions. The full affiliation of each group can be seen in their papers in this volume.

The introduction of fully fledged question answering and image retrieval tracks had a big impact on participation in CLEF 2004, not just with respect to the numbers but also regarding the skills and expertise involved. The popularity of question answering has meant that a growing number of participants have a natural language processing background while the image retrieval tasks have brought in groups with experience in new areas including image processing and medical informatics – making CLEF an increasingly multidisciplinary forum.

Figure 1 shows how the focus of CLEF has shifted and diversified over the years.

Acad. Sciences/ITC-irst (BG/IT)
CEA/LIC2M (FR) *
CLIPS-IMAG/IPAL-CNRS (FR/SG)*
Clairvoyance Corp. (US) **
Daedalus/Madrid Universities (ES) *
DFKI (DE) *
Dublin City U. (IE)
Hummingbird (CA) ***
ILC-CNR/U.Pisa (IT)
Imperial College London (UK)
INAOE (MX)
IRIT-Toulouse (FR) **
ITC-irst (IT) ****
Johns Hopkins U. (US) ****
Linguateca SINTEF (NO)
LMSI-CNRS (FR)
National Research Council (CA)
National Taiwan U. (TW) ***
KIDS - NCTU/ISU (TW)
Ricoh (JP) *
SICS/Connexor (SV/FI) ***
SUNY at Buffalo (US) *
Thomson Legal & Reg. (US)***
U.Alicante (ES) ***
U.Amsterdam (NL) ***
U.Chicago (US)
U.Evora (PT)
U.Edinburgh (UK)
U.Glasgow (UK)
U.Hagen (DE) *
U.Helsinki (FI)
U.Hildesheim (DE) **
U.Hospitals Geneva/LITH (CH)
U.Jaen (ES) ***
U.La Coruna (ES) **
U.Limerick (IE) *
U.Lisbon (PT)
U.Maryland (US) ****
U. Michigan (US)
U.Montreal (CA) ****
U.Neuchâtel (CH) ***
U.Oregon (US)
U.Oviedo (ES) *
U.Padova (IT) **
U.Salamanca (ES) **
U.Sheffield (UK) ****
U.Stockholm/SICS (SV)
U.Surugadai/NII/NTU (JP/TW) *
U.Tech Aachen – Comp.Sci (DE)
U.Tech Aachen – Medicine (DE)
U.Tilburg/U.Maastricht (NL)
U.Twente/CWI (NL) ***
UC Berkeley (US) ****
UNED (ES) ***
UP Catalunya (ES)

Table 3: CLEF2004 Participating Groups

4. Working Notes and Workshop

The Working Notes provide a first description of the different experiments made by this year's participants. They consist of two volumes. Volume I contains 80 papers and is divided into eight sections, mainly corresponding to the CLEF 2004 tracks. The results of the ad hoc and domain specific tracks are reported in the first three sections: Cross-language and More contains papers describing multilingual and bilingual experiments, which may also include details on monolingual work, whereas the second section contains papers that focus on Monolingual Experiments only. Section 3 provides reports on monolingual and bilingual system testing on the GIRT social science database. Each of the next four sections – dedicated to the results of the iCLEF, QA@CLEF, ImageCLEF and CL-SDR tracks - begins with a overview by the track coordinators followed by papers describing the experiments of the participating groups.

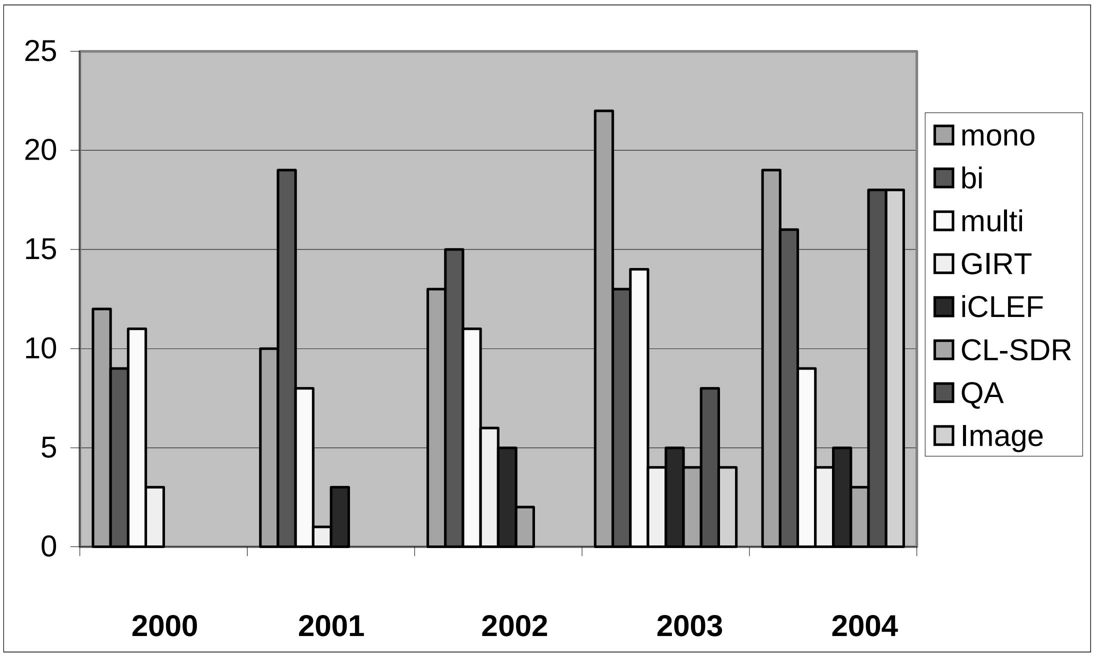

Figure 1 CLEF 200 – 2004 - numbers of participants per track
(in 2002 the GIRT track also included the Amaryllis database)

The final section contains the outline of one of the two invited talks at the Workshop plus two papers discussing issues that affect multilingual system testing and evaluation. Volume II contains three appendices. Appendix A gives a list of the characteristics of all runs for the ad hoc and GIRT tracks together with overview graphs for the different tasks and individual statistics for each run. Appendices B and C contain run statistics for the question answering and the image retrieval tracks, respectively.

CLEF aims at creating a strong multilingual information access 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 discussions of the results of individual tracks and intentions for the future. The invited talks mentioned above will consist of a discussion on building CLIR applications and a report on the recent activities of the NTCIR evaluation initiative for Asian languages. The final sessions will include discussions on ideas for future tracks and a panel on new directions for CLEF. 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 for 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

CLEF is organised 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 contributed to the coordination of the 2004 campaigns and the organisation of the Workshop. Without their assistance, this initiative would be impossible. A list of the principle institutions involved is given in the following pages. Here below, let me thank the main track coordinators:

- Martin Braschler, Eurospider Information Technologies, Switzerland, for the Ad hoc Track
- Michael Kluck, IZ-Bonn, Germany, for the GIRT track
- Julio Gonzalo, LSI-UNED, Madrid, Spain, and Douglas W. Oard, U. Maryland, USA, for iCLEF
- Bernardo Magnini, ITC-irst, Trento, Italy, for QA@CLEF
- Paul Clough, U. Sheffield, UK , and Henning Müller, U. Hospitals of Geneva, Switzerland, for ImageCLEF
- Marcello Federico, ITC-irst, Trento, Italy, and Gareth Jones, DCU, Ireland, for CL-SDR

In addition, I must thank colleagues from the Natural Language Processing Lab, Department of Computer Science and Information Engineering, National Taiwan University and the National Institute of Informatics, Tokyo, for preparing Chinese and Japanese topics. I also thank the European Language Resources Association for its sponsorship of the workshop.

I should also like to express my gratitude to 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;
- 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 Sozialwissenschaften, Bonn, for the GIRT database.
- Hypersystems Srl, Torino and La Stampa, for the Italian newspaper data.
- Agencia EFE S.A. for the Spanish newswire data.
- NRC Handelsblad, Algemeen Dagblad and PCM Landelijke dagbladen/Het Parool for the Dutch newspaper data.
- Aamulehti Oyj for the Finnish newspaper documents
- Tidningarnas Telegrambyrå for the Swedish newspapers

- The Herald 1995, SMG Newspapers, for the British English newspaper data
- Público and Linguateca for the Portuguese newspaper collection
- Schweizerische Depeschenagentur, Switzerland, for the French, German and Italian Swiss news agency data.
- Russika-Izvestia for the Russian collection
- St Andrews University Library for the image collection
- Radiology Dept. University Hospitals, Geneva, Switzerland for the Medical Images Database
- NIST for access to the TREC-8 and TREC-9 SDR transcripts.

Without their help, this evaluation activity would be impossible.

Last but not least, I should like to thank both Francesca Borri in Pisa and Natasha Bishop, UKOLN, University of Bath, for their assistance in the local organisation of the CLEF 2004 Workshop.

Carol Peters
CLEF Coordinator
September 2004

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 2004 campaign:

- 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 and Information Systems, University of Limerick, Ireland
- Department of Information Studies, University of Sheffield, UK
- Department of Information Studies, University of Tampere, Finland
- Eurospider Information Technology AG, Zürich, Switzerland
- 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 Sistemás Informáticos, Universidad Nacional de l'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
- School of Computing, Dublin City University, Ireland
- University Hospitals of Geneva, Switzerland

CLEF Steering Committee

Maristella Agosti, University of Padova, Italy
Eija Airio, University of Tampere, Finland
Martin Braschler, Eurospider Information Technologies, Zurich, Switzerland
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
Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
Marcello Federico, ITC-irst, Trento, Italy
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, Informationszentrum Sozialwissenschaften Bonn, Germany
Natalia Loukachevitch, Moscow State University, Russia
Bernardo Magnini, ITC-irst, Trento, Italy
Paul McNamee, Johns Hopkins University, USA
Henning Müller, University Hospitals of Geneva, Switzerland
Douglas W. Oard, University of Maryland, USA
Maarten de Rijke, University of Amsterdam, Netherlands
Jacques Savoy, University of Neuchâtel, Switzerland
Peter Schäuble, Eurospider Information Technologies, Switzerland
Richard Sutcliffe, University of Limerick, Ireland
Hans Uszkoreit, German Research Center for Artificial Intelligence (DFKI), Germany
José Luis Vicedo, University of Alicante, Spain
Ellen Voorhees, National Institute of Standards and Technology, USA
Christa Womser-Hacker, University of Hildesheim, Germany