What happened in CLEF 2009

Introduction to the Working Notes

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

The objective of the Cross Language Evaluation Forum[1] is to promote research in the field of multilingual system development. This is done through the organisation of annual evaluation campaigns in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval (IR) are offered. The intention is 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. This has been achieved by offering an increasingly complex and varied set of evaluation tasks over the years. The aim is not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems.

These Working Notes contain descriptions of the experiments conducted within CLEF 2009 – the tenth in a series of annual system evaluation campaigns. The results of the experiments will be presented and discussed in the CLEF 2009 Workshop, 30 September – 2 October, Corfu, Greece. 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 2009 Proceedings, to be published by Springer in their Lecture Notes for Computer Science series.

Since CLEF 2005, the Working Notes are published in electronic format only and are distributed to participants at the Workshop on a memory stick together with a printed volume of Extended Abstracts. The Working Notes for all the ten CLEF campaigns can be found in electronic form on the CLEF website at www.clef-campaign.org.

Both the 2009 Working Notes and Book of Abstracts are divided into ten sections, corresponding to the eight main evaluation tracks, the experimental pilot task, and another evaluation initiative using CLEF data: Morpho Challenge 2009. Appendices are also included containing run statistics for the Ad Hoc and Grid@CLEF tracks, and a list of all participating institutions.

The main features of the 2009 campaign are briefly outlined in the following sections in order to provide the necessary background to the experiments reported in the rest of the Working Notes.

1. Tracks and Tasks in CLEF 2009

CLEF 2009 offered eight main tracks designed to evaluate the performance of systems for:

- multilingual textual document retrieval (Ad Hoc)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)
- multilingual information filtering (INFILE@CLEF)
- cross-language video retrieval (VideoCLEF)
- intellectual property (CLEF-IP) – New this year
- log file analysis (LogCLEF) – New this year

An experimental pilot task was also offered:

- Grid Experiments (Grid@CLEF)

In addition, Morpho Challenge 2009 was organized in collaboration with CLEF as part of the EU Network of Excellence Pascal Challenge Program[2]. The Morpho Challenge participants will meet separately, before the main CLEF workshop on the morning of Wednesday 30 September, to discuss their results.

[1] Since the beginning of 2008, CLEF is included in the activities of the TrebleCLEF Coordination Action, funded by the Seventh Framework Programme of the European Commission. For information on TrebleCLEF, see www.trebleclef.eu.
[2] Morpho Challenge is part of the EU Network of Excellence Pascal: http://www.cis.hut.fi/morphochallenge2009/

Here below we give a brief overview of the various activities.

Multilingual Textual Document Retrieval (Ad Hoc): The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. From 2000 - 2007, the track exclusively used collections of European newspaper and news agency documents. Last year the focus of the track was considerably widened: we introduced very different document collections, a non-European target language, and an information retrieval (IR) task designed to attract participation from groups interested in natural language processing (NLP). The 2009 Ad Hoc track was to a large extent a repetition of last year's track, with the same three tasks: Tel@CLEF, Persian@CLEF, and Robust-WSD. An important objective has been to create good reusable test collections for each of them The track was thus structured in three distinct streams. The first task offered monolingual and cross-language search on library catalog records and was organized in collaboration with The European Library (TEL)[3]. The second task resembled the ad hoc retrieval tasks of previous years but this time the target collection was a Persian newspaper corpora. The third task was the robust activity which used word sense disambiguated (WSD) data. The track was coordinated jointly by ISTI-CNR and Padua University, Italy; the University of the Basque Country, Spain; with the collaboration of the Database Research Group, University of Tehran, Iran.

Interactive Cross-Language Retrieval (iCLEF): In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language. Since 2006, iCLEF has based its experiments on Flickr, a large-scale, web-based image database where image annotations constitute a naturally multilingual folksonomy. In an attempt to encourage greater participation in user-orientated experiments, a new task was designed for 2008 and has had a continuation in 2009. The main novelty has been to focus experiments on a shared analysis of a large search log, generated by iCLEF participants from a single search interface provided by the iCLEF organizers. The focus is, therefore, on search log analysis rather than on system design. The idea is to study the behaviour of users in an (almost) naturalistic search scenario, having a much larger data set than in previous iCLEF campaigns. The track was coordinated by UNED, Madrid, Spain; Sheffield University, UK; Swedish Institute of Computer Science, Sweden.

Multilingual Question Answering (QA@CLEF): This track has offered monolingual and cross-language question answering tasks since 2003. QA@CLEF 2009 proposed three exercises: ResPubliQA, QAST and GikiCLEF:

- ResPubliQA: The hypothetical user considered for this exercise is a person close to the law domain interested in making inquiries on European legislation. Given a pool of 500 independent natural language questions, systems must return the passage that answers each question (not the exact answer) from the JRC-Acquis collection of EU parliamentary documentation. Both questions and documents are translated and aligned for a subset of languages. Participating systems could perform the task in Basque, Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish.
- QAST: The aim of the third QAST exercise was to evaluate QA technology in a real multilingual speech scenario in which written and oral questions (factual and definitional) in different languages are formulated against a set of manually and automatically transcribed audio recordings related to speech events in those languages. The scenario proposed was the European Parliament sessions in English, Spanish and French.
- GikiCLEF: Following the previous GikiP pilot at GeoCLEF 2008, the task focused on open list questions over Wikipedia that require geographic reasoning, complex information extraction, and cross-lingual processing, for collections in Bulgarian, Dutch, English, German, Italian, Norwegian (both Bokmål and Nynorsk), Portuguese and Romanian or Spanish.

The track was organized by a number of institutions (one for each target language), and jointly coordinated by CELCT, Trento, Italy, and UNED, Madrid, Spain.

Cross-Language Retrieval in Image Collections (ImageCLEF): This track evaluated retrieval from visual collections; both text and visual retrieval techniques were employed. A number of challenging tasks were offered:

- multilingual ad-hoc retrieval from a photo collection concentrating on diversity in the results;
- a photographic annotation task using a simple ontology;
- retrieval from a large scale, heterogeneous collection of Wikipedia images with user-generated textual metadata, and queries in several languages;
- medical image retrieval (with visual, semantic and mixed topics in several languages);
- medical image annotation;
- detection of semantic categories from robotic images (non-annotated collection, concepts to be detected).

[3] See http://www.theeuropeanlibrary.org/

A large number of organisations have been involved in the complex coordination of these tasks. They include: Sheffield University, UK; University of Applied Sciences Western Switzerland; Oregon Health and Science University, USA; University of Geneva, Switzerland; CWI, The Netherlands; IDIAP, Switzerland.
The ImageCLEF track has organised a separate one day workshop on visual information retrieval evaluation in collaboration with the Theseus project; this will be held immediately before the main CLEF workshop, on 29 September at the University of Corfu.

Multilingual Information Filtering (INFILE@CLEF): INFILE (INformation, FILtering & Evaluation) is a cross-language adaptive filtering evaluation track sponsored by the French National Research Agency. INFILE has extended the last filtering track of TREC 2002 as follows. It uses a corpus of 100,000 Agence France Press comparable newswires for Arabic, English and French; evaluation is performed using an automatic querying of test systems with a simulated user feedback. Each system can use the feedback at any time to increase performance. The track has been coordinated by the Evaluation and Language resources Distribution Agency (ELDA), France; University of Lille, France; and CEA LIST, France.

Cross-Language Video Retrieval (VideoCLEF): VideoCLEF 2009 is dedicated to developing and evaluating tasks involving access to video content in a multilingual environment. Participants were provided with a corpus of video data (Dutch-language television, predominantly documentaries) accompanied by speech recognition transcripts. In 2009, there were three tasks: "Subject Classification", which involved automatically tagging videos with subject labels; "Affect", which involved classifying videos according to characteristics beyond their semantic content; ``Finding Related Resources Across Languages", which involved linking video to material on the same subject in a different language. The track was jointly coordinated by Delft University of Technology and Dublin City University, Ireland.

Intellectual Property (CLEF-IP): This was the first year for the CLEF-IP track. The purpose of the track was twofold: to encourage and facilitate research in the area of patent retrieval by providing a large clean data set for experimentation; to create a large test collection of patents in the three main European languages for the evaluation of cross-lingual information access. The track focused on the task of prior art search. A large test collection for evaluation purposes was created by exploiting patent citations. The collection consists of a corpus of 1,9 million patent documents and 10,000 topics with an average of 6 relevance assessments per topic.

Log file analysis (LogCLEF): LogCLEF is an evaluation initiative for the analysis of queries and other logged activities as expression of user behaviour. The goal is the analysis and classification of queries in order to understand search behaviour in multilingual contexts and ultimately to improve search systems. The track used log data from the files of The European Library.

Grid Experiments (Grid@CLEF): This experimental pilot is planned as a long term activity with the aim of: looking at differences across a wide set of languages; identifying best practices for each language; helping other countries to develop their expertise in the IR field and create IR groups. Participants had to conduct experiments according to the CIRCO (Coordinated Information Retrieval Components Orchestration) protocol, an XML-based framework which allows for a distributed, loosely-coupled, and asynchronous experimental evaluation of Information Retrieval (IR) systems. The track is coordinated jointly by University of Padua, Italy, National Institute of Standards and Technology, USA.

Unsupervised Morpheme Analysis (Morpho Challenge): Morpheme analysis is particularly useful in speech recognition, information retrieval and machine translation for morphologically rich languages where the amount of different word forms is very large. In Morpho Challenge 2009 unsupervised algorithms that provide morpheme analyses for words in different languages were evaluated in various practical applications. The evaluations consisted of: 1) a comparison to grammatical morphemes,2) using morphemes instead of words in information retrieval tasks, and 3) combining morpheme and word based systems in statistical machine translation tasks. The evaluation languages in 2009 were: Finnish, Turkish, German, English and Arabic. The track was coordinated by Helsinki University of Technology and Cambridge University Engineering Department.

Details on the technical infrastructure and the organisation of all these tracks can be found in the track overview reports in this volume, collocated at the beginning of the relevant sections. Figure 1 shows the ratio of participants per track.

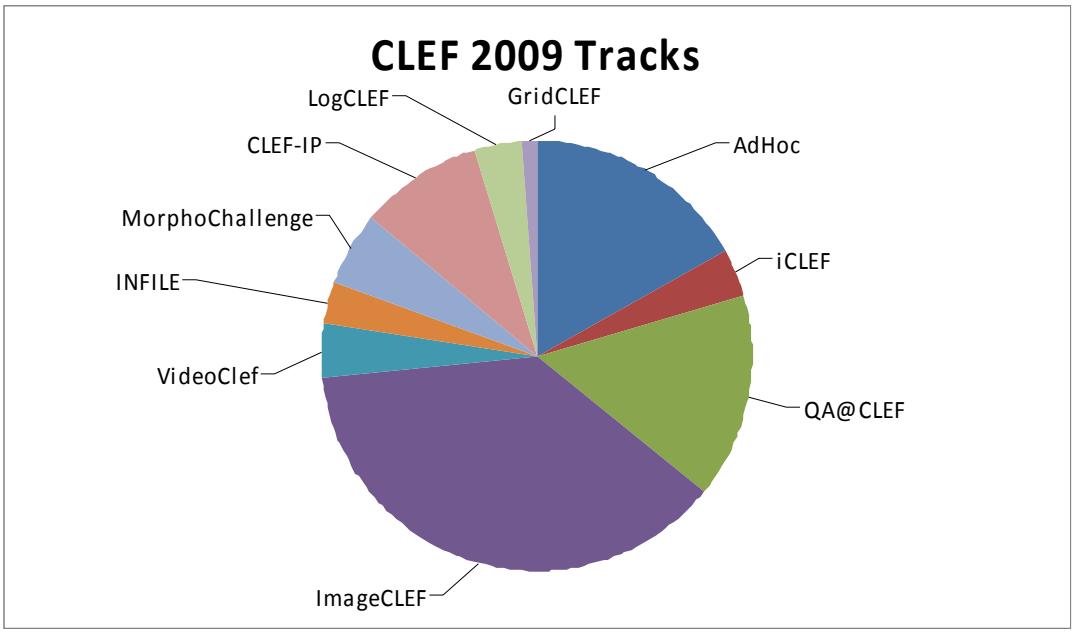

Figure 1. CLEF 2009: Participation Track by Track

2. Test Collections

The CLEF test collections are made up of documents, topics and relevance assessments. The topics are created to simulate particular information needs from which the systems derive the queries to search the document collections. System performance is evaluated by judging the results retrieved in response to a topic with respect to their relevance, and computing the relevant measures, depending on the methodology adopted by the track. The document sets that have been used to build the test collections in CLEF 2009 included:

- A subset of the CLEF multilingual corpus of news documents in 14 European languages (Ad Hoc WSD-Robust task)
- Hamshahri Persian newspaper corpus (Ad Hoc Persian task)
- Library catalog records in English, French, German plus log files provided by The European Library (Ad Hoc TEL task and LogCLEF)
- Flickr web-based image database (iCLEF)
- ResPubliQA document collection, a subset of the JRC Acquis corpus of European legislation (QAatCLEF: ResPubliQA)
- Transcripts of European parliamentary sessions in English and Spanish, and French news broadcasts (QAatCLEF: QAST)
- BELGAPICTURE image collection (ImageCLEFPhoto)
- Multilingual collections of Wikipedia documents and images (ImageCLEFwiki)
- Articles and images from Radiology and Radiography; IRMA collection for medical image annotation (ImageCLEFmed and medAnnotation)
- Dutch and English documentary television programs (VideoCLEF)
- Agence France Press (AFP) comparable newswire stories in Arabic, French and English (INFILE)
- Patent documents in English, French and German from the European Patent Office (CLEF-IP)

Acknowledgements of the valuable contribution of the data providers is given at the end of this paper.

3. CLEF & TrebleCLEF

CLEF is organized mainly through the voluntary efforts of many different institutions and research groups. Section 1 gives the groups responsible for the coordination of this year's tracks. A full list of the people and groups involved in the organization of CLEF2009 is given at the end of this paper. However, the central coordination has always received some support from the EU IST programme under the unit for Digital Libraries and Technology Enhanced Learning, mainly within the framework of the DELOS Network of Excellence. CLEF 2008 and 2009 are organized under the auspices of TrebleCLEF, a Coordination Action of the Seventh Framework Programme, Theme ICT 1-4-1.

TrebleCLEF is building on and extending the results already achieved by CLEF. The objective is to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access and to promote a dissemination action in the relevant application communities.
The aim is to

- Provide applications that need multilingual search solutions with the possibility to identify the technology which is most appropriate
- Assist technology providers to develop competitive multilingual search solutions.

Information on the activities of TrebleCLEF can be found on the project website.

4. Technical Infrastructure

TrebleCLEF supports a data curation approach within CLEF as an extension to the traditional methodology in order to better manage, preserve, interpret and enrich the scientific data produced, and to effectively promote the transfer of knowledge. The current approach to experimental evaluation is mainly focused on creating comparable experiments and evaluating their performance whereas researchers would also greatly benefit from an integrated vision of the scientific data produced, together with analyses and interpretations, and from the possibility of keeping, re-using, and enriching them with further information. The way in which experimental results are managed, made accessible, exchanged, visualized, interpreted, enriched and referenced is an integral part of the process of knowledge transfer and sharing towards relevant application communities.

The University of Padua has thus developed DIRECT: Distributed Information Retrieval Evaluation Campaign Tool[4], a digital library system for managing the scientific data and information resources produced during an evaluation campaign. A preliminary version of DIRECT was introduced into CLEF in 2005 and subsequently tested and developed in the CLEF 2006 and 2007 campaigns. It has been further developed under TrebleCLEF.

DIRECT currently manages the technical infrastructure for several of the CLEF tracks and tasks: Ad Hoc, ImageCLEFphoto, GridCLEF, providing procedures to handle:

- the track set-up, harvesting of documents, management of the registration of participants to tracks;
- the submission of experiments, collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- the provision of common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- the provision of common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses.

DIRECT is designed and implemented by Giorgio Di Nunzio, Nicola Ferro and Marco Dussin.

5. Participation

Researchers from 117 different academic and industrial institutions submitted runs in CLEF 2009: 81 from Europe, 18 from N.America; 16 from Asia, 1 from S.America and 1 from Africa. The breakdown of participation of groups per track is as follows: Ad Hoc 28(26); iCLEF 6(6); QA@CLEF 25(29); ImageCLEF 62(42); INFILE 5(1); VideoCLEF 7(5); CLEF-IP 15; LogCLEF 6; Grid@CLEF 2; MorphoChallenge 9(6).Last years figures are given between brackets where applicable. A list of these institutions and indications of the tracks in which they participated is included in an Appendix to these Working Notes. Figure 2 shows the trend in participation over the years and Figure 3 shows the shift in focus as new tracks are added.

As can be seen, the number of groups participating in the Ad Hoc, iCLEF, QA and VideoCLEF tracks is almost the same as last year, there has been a rise of interest in INFILE and participation in the two new tracks (LogCLEF and CLEF-IP) is encouraging. The most popular track is without doubt ImageCLEF which, with a notable increase from the 42 groups of last year, is now dominating the scene. This gives some cause for reflection as ImageCLEF is the track least concerned with multilinguality. However, recognising that multilingual information access is a truly multidisciplinary domain, one of the main objectives of CLEF has always been to create a forum where researchers from a wide range of areas can get together. CLEF is perhaps one of the few platforms where groups working in many different areas (e.g. Information Retrieval, Natural Language Processing, Image Processing, Speech Recognition, Log Analysis, etc., etc.) have a chance to see what others are doing, and discuss and compare ideas.

[4] http//direct.dei.unipd.it/

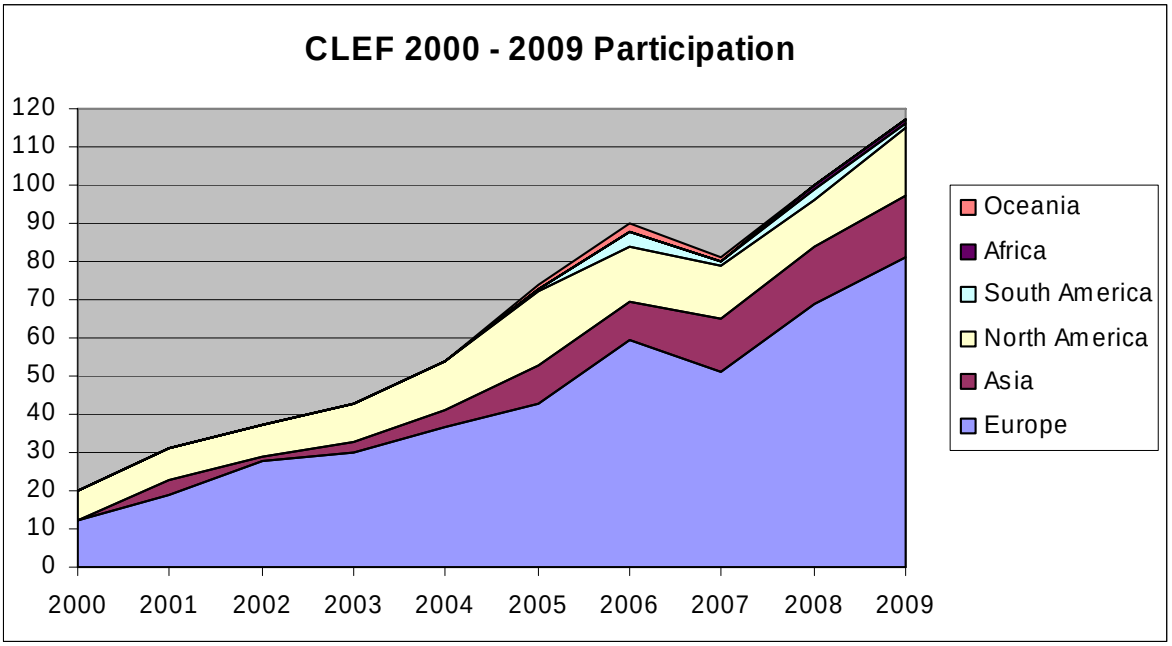

Figure 2. CLEF 2000 – 2009: Participation

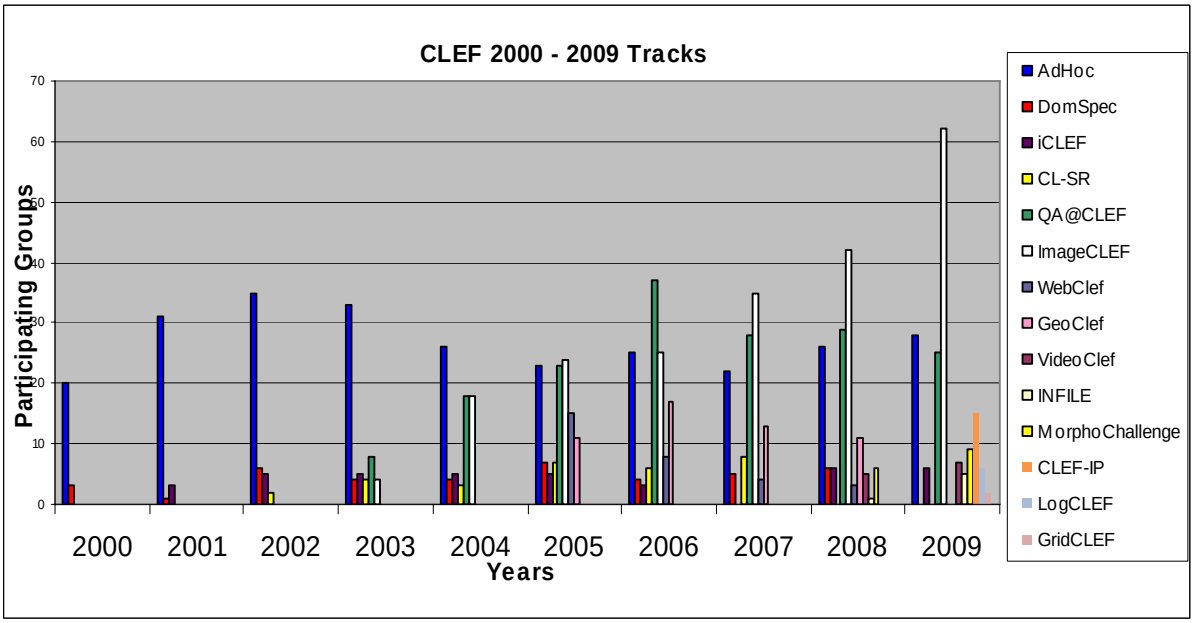

Figure 3. CLEF 2000 – 2009: Participation per Track

6. Workshop

As has been stated, CLEF aims at creating a strong CLIR/MLIA 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 and parallel 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 the future. 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 are posted on the CLEF website at http://www.clef-campaign.org.

7. The Future of CLEF

CLEF has been running for almost ten years now with the main goal of sustaining the growth of excellence in language processing and multilingual information access (MLIA) across language boundaries within the global context of the multilingual Web. Over the years, strongly motivated by the need to promote the study and utilisation of languages other than English on the Internet, a core network of research institutions involved with CLEF, with some support for the central coordination mainly from the DELOS Network of Excellence for Digital Libraries, has produced the following significant results:

- Creation of a very active multidisciplinary international research community, with strong interactions with both TREC and NTCIR including coordination of schedules and activities;
- Investigation of core issues in MLIA which enable effective transfer over language boundaries, including the development of multiple language processing tools (e.g. stemmers, word decompounders, part-of-speech taggers); creation of linguistic resources (e.g. multilingual dictionaries and corpora); implementation of appropriate cross-language retrieval models and algorithms for different tasks and languages;
- Creation of important reusable test collections and resources in diverse media for a large number of European languages, representative of the major European language typologies;
- Significant and quantifiable improvements in the performance of MLIA systems;

However, since CLEF began the associated technologies, services and users of multilingual IR systems have been in continual evolution, with many new factors and trends influencing the field. For example, the growth of the Internet has been exponential with respect to the number of users and languages used regularly for global information dissemination. The expectations and habits of users are constantly changing, together with the ways in which they interact with content and services, often creating new and original ways of exploiting them. Language barriers are no longer seen as inviolable and there is a growing dissatisfaction with the technologies currently available to overcome them.

This constantly evolving scenario poses challenges to the research community which must react to these new trends and emerging needs. CLEF initially assumed a user model reflecting simple information seeking behaviour: the retrieval of a list of relevant items in response to a single query that could then be used for further consultation in various languages and media types. This simple scenario of user interaction has allowed researchers to focus their attention on studying core technical issues for CLIR systems and associated components.

If we are to continue advancing the state-of-the-art in multilingual information access technologies, we now need to rethink and update this user model. We have to study and evaluate multilingual issues from a communicative perspective rather than a purely retrieval one. We need to examine the interactions between four main entities: users, their tasks, languages, and content in order to understand how these factors impact on the design and development of MLIA systems. It is not sufficient to successfully cross the language boundary, results must be retrieved in a form that is interpretable and reusable. Future cross-language system evaluation campaigns must activate new forms of experimental evaluation - laboratory and interactive – in order to foster the development of MLIA systems more adherent to the new user needs. We need a deeper understanding of the interaction between multicultural and information proactive users, multilingual content, language-dependent tasks, and the enabling technologies consisting of MLIA systems and their components.

At the same time, benchmarking efforts must prove their usefulness for industrial take-up; evaluation initiatives risk being seen as irrelevant for system developers if the data they investigate are not of realistic scale and if the use cases and scenarios tested do not appear valid.

Future editions of CLEF should thus introduce a new series of evaluation cycles which move beyond the current set-up, impacting on:

- Methodology definition: evolution of the current evaluation paradigm, developing new models and metrics to describe the needs and behavior of the new multicultural and multi-tasking users;
- System building: driving the development of MLIA systems and assessing their conformity with respect to the newly identified user needs, tasks, and models;
- Results assessment: measuring all aspects of system & component performance including response times, usability, and user satisfaction;
- Community building: promoting the creation of a multidisciplinary community of researchers which goes beyond the existing CLEF community by building bridges to other relevant research domains such as the MT,

information science and user studies sectors, and to application communities, such as the enterprise search, legal, patent, educational, cultural heritage and infotainment areas;
- Validation of technology: providing a reasonably comprehensive typology of use cases and usage scenarios for multilingual search, validated through user studies, to enable reuse of appropriate resources and to enable common evaluation schemes;
- Technology transfer: guaranteeing that the results obtained are demonstrated as useful for industrial deployment.

Achieving this goal will require further synergy between various research communities including machine translation, information retrieval, question answering, information extraction, and representatives from end user groups. If this programme is to be implemented, it is clear that CLEF – or any similar evaluation initiative – needs a solid underlying management and coordination structure We feel that the ten year milestone may be the right point to pause for a moment to rethink carefully the best way in which to continue to ensure that the programme of activities is viable, consistent and coherent and that CLEF can successfully scale up and embrace new communities and technological paradigms.

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 2009 campaigns. A list of the main institutions involved is given in the following pages. Here below, let me thank just some of the people responsible for the coordination of the different tracks. My apologies to all those I have not managed to mention:

- Abolfazl AleAhmad, Hadi Amiri, Eneko Agirre, Giorgio Di Nunzio, Nicola Ferro, Nicolas Moreau, Arantxa Otegi and Vivien Petras for the Ad Hoc Track
- Paul Clough, Julio Gonzalo and Jussi Karlgren for iCLEF
- Iñaki Alegria, Davide Buscaldi, Luís Miguel Cabral, Pere R. Comas, Corina Forascu, Pamela Forner, Olivier Galibert, Danilo Giampiccolo, Nicolas Moreau, Djamel Mostefa, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Sophie Rosset, Paolo Rosso, Diana Santos, Richard Sutcliff and Jordi Turmo for QA@CLEF
- Brian Bakke, Steven Bedrick, Barbara Caputo, Paul Clough, Peter Dunker, Thomas Deselaers, Thomas Deserno, Ivan Eggel, Mark Oliver Güld, William Hersh, Patric Jensfelt, Charles E. Kahn Jr., Jana Kludas, Jayashree Kalpathy–Cramer, Henning Müller, Stefanie Nowak, Monica Lestari Paramita, Andrzej Pronobis, Saïd Radhouani, Mark Sanderson, Tatiana Tommasi, Theodora Tsikrika and Petra Welter for ImageCLEF
- Romaric Besançon, Stéphane Chaudiron, Khalid Choukri, Meriama Laïb, Djamel Mostefa and Ismaïl Timimi for INFILE
- Gareth J.F. Jones, Martha Larson and Eamonn Newman for VideoCLEF
- Florina Piroi, Giovanna Roda, John Tait and Veronika Zenz for CLEF-IP
- Maristella Agosti, Giorgio Di Nunzio, Christine Doran, Inderjeet Mani, Thomas Mandl, Julia Maria Schulz and Alexander Yeh for LogCLEF
- Nicola Ferro and Donna Harman for GridCLEF
- Graeme W. Blackwood, William Byrne Mikko Kurimo, Ville T. Turunen and Sami Virpioja for MorphoChallenge at CLEF
- Marco Duissin, Giorgio Di Nunzio and Nicola Ferro for developing and managing the DIRECT infrastructure.

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. Without their contribution, this evaluation activity would be impossible.

Finally, I should like to express my gratitude to Francesca Borri and Alessandro Nardi in Pisa and Giannis Tsakonas in Corfu for their assistance in the organisation of the CLEF 2009 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 2009 campaign:

- Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
- Berlin School of Library and Information Science, Humboldt-Universität zu Berlin
- Business Information Systems, University of Applied Sciences Western Switzerland, Sierre, Switzerland
- CEA LIST, France
- Center for Autonomous Systems, Royal Institute of Technology, Sweden
- Center for Evaluation of Language and Communication Technologies (CELCT), Italy
- Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands
- Computer Science Department, University of the Basque Country, Spain
- Computer Vision and Multimedia Lab, University of Geneva, Switzerland
- Database Research Group, University of Tehran, Iran
- Department of Computer Science & Information Systems, University of Limerick, Ireland
- Department of Information Engineering, University of Padua, Italy
- Department of Information Science, University of Hildesheim, Germany
- Department of Information Studies, University of Sheffield, UK
- Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, USA
- Department of Medical Informatics, University Hospitals and University of Geneva, Switzerland
- Department of Medical Informatics, Aachen University of Technology (RWTH), Germany
- Evaluations and Language Resources Distribution Agency Sarl, Paris, France
- Fraunhofer Institute for Digital Media Technology (IDMT), Ilmenau, Germany
- GERiiCO, Université de Lille, France
- Idiap Research Institute, Switzerland
- Information Retrieval Facility (IRF), Austria
- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Orsay, France
- Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguateca, SINTEF ICT, Norway
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria
- Matrixware Information Services, Austria
- Mediamatics, Delft University of Technology, The Netherlands
- Mitre Corporation, USA
- National Institute of Standards and Technology, Gaithersburg MD, USA
- NLE Lab., Universidad Politènica de Valencia, Spain
- Research Institute for Artificial Intelligence, Romanian Academy, Romania
- Romanian Institute for Computer Science, Romania
- Royal Institute of technology (KTH), Stockholm, Sweden
- School of Computing, Dublin City University, Ireland
- Swedish Institute of Computer Science, Sweden
- TALP Research Center, Universitat Politécnica de Catalunya, Barcelona, Spain.

CLEF Steering Committee

- Maristella Agosti, University of Padova, Italy
- Martin Braschler, Zurich University of Applied Sciences Winterhur, 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
- Thomas Deselaers, RWTH Aachen University, Germany
- Giorgio Di Nunzio, University of Padova, Italy
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, ITC-irst, Trento, Italy
- Nicola Ferro, University of Padova, 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, The 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 of Applied Sciences Western Switzerland, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Anselmo Peñas, LSI-UNED, Madrid, Spain
- Vivien Petras, Humboldt University, Berlin, Germany
- Maarten de Rijke, University of Amsterdam, The Netherlands
- Diana Santos, Linguateca, Sintef, Oslo, Norway
- Jacques Savoy, University of Neuchatel, Switzerland
- Peter Schäuble, Eurospider Information Technologies, Switzerland
- Richard Sutcliffe, University of Limerick, Ireland
- Denis Teyssou, Agence France-Presse, Paris, France
- 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

Data Providers

The support of all the data providers and copyright holders that have contributed to the creation of the CLEF test collections over the years is gratefully acknowledged, and in particular:

- The Los Angeles Times, for the American-English newspaper collection.
- SMG Newspapers (The Herald) for the British-English newspaper collection.
- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French newspaper collection.
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections.
- Hypersystems Srl, Torino and La Stampa, for the Italian newspaper data.
- Agencia EFE S.A. for the Spanish news agency 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.
- Hamshahri newspaper and DBRG, Univ. Tehran, for the Persian newspaper data.
- Público, Portugal, and Linguateca for the Portuguese (PT) newspaper collection.
- Folha de São Paulo, 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 & 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, Novinar OD, Sofia and the BulTreeBank Project, Linguistic Modelling Laboratory, IPP, Bulgarian Acad. Sci, for the Bulgarian newspaper documents
- Mafra a.s. and Lidové Noviny a.s. for the Czech newspaper data
- Usurbilgo Udala, Basque Country, Spain, for the Egunkaria, Basque newspaper documents
- The European Commission – Joint Research Centre for the JRC Acquis Parallel corpus of European legislation in many languages.
- AFP Agence France Presse for the English, French and Arabic newswire data used in the INFILE track
- The British Library, Bibliothèque Nationale de France and the Austrian National Library for the library catalog records forming part of The European Library (TEL)
- The European Library (TEL) for use of TEL log files
- Tumba! web search engine of the Faculdade de Ciências da Universidade de Lisboa (FCUL), Portugal, for logfile querying
- InformationsZentrum Sozialwissen-schaften, Bonn, for the GIRT social science database.
- SocioNet system for the Russian Social Science Corpora
- Institute of Scientific Information for Social Sciences of the Russian Academy of Science (ISISS RAS) for the ISISS database
- Aachen University of Technology (RWTH), Germany, for the IRMA annotated medical images.
- Mallinkrodt Institue of Radiology for permission to use their nuclear medicine teaching file.
- Radiological Society of North America for the images of the Radiology and Radiographics journals.
- Lung Image Database Consortium (LIDC) for their database of lung nodules.
- University of Basel's Pathopic project for their Pathology teaching file
- Belga Press Agency, Belgium, for BELGAPICTURE image collection
- LIACS Medialab, Leiden University, The Netherlands & Fraunhofer IDMT, Ilmenau, Germany for the use of the MIRFLICKR 25000 Image collection
- Wikipedia for the use of the Wikipedia image collection.

- USC Shoah Foundation Institute, and IBM (English) and The Johns Hopkins University Center for Language and Speech Processing (Czech) for the speech transcription
- ELDA for the use of the ESTER Corpus: Manual and automatic transcripts of French broadcast news
- ELDA for the use of EPPS 2005/2006 ES & EN Corpora: Manual and automatic transcriptions of European Parliament Plenary Sessions in Spanish and English
- Matrixware Information Services GmbH for the use of a collection of patent documents in English, French and German from the European Patent Office
- The Institute of Sound and Vision, The Netherlands, for the English/Dutch videos, the University of Twente for the speech transcriptions, and Dublin City University for the shot segmentation.

Without their contribution, this evaluation activity would be impossible.