# Welcome to CLEF 2006

Carol Peters

ISTI-CNR Pisa, Italy

# Cross-Language System Evaluation


## 10 years of activity

- CLIR track at TREC (1997-1999)
- CLEF 2001 & 2000 - sponsored by DELOS Network of Excellence (5FP) and US National Institute of Standards and technology
- CLEF 2002 & 2003 - IST-2000-31002
- CLEF 2004, 2005 & 2006 again sponsored by DELOS Network of Excellence

**plus**

# CLEF Coordination


CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa
The following Institutions are contributing to the organisation of the different tracks of the CLEF 2006 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, U. Maryland, USA
- Dept. of Computer Science, U. Indonesia
- Depts. of Computer Science & Medical Informatics, RWTH Aachen U., Germany
- Dept. of Computer Science and Information Systems, U. Limerick, Ireland
- Dept. of Computer Science and Information Engineering, National U. Taiwan
- Dept. of Information Engineering, U. Padua, Italy
- Dept. of Information Sci, U. Hildesheim, Germany
- Dept. of Information Studies, U. Sheffield, UK
- Evaluations and Language Resources Distribution Agency Sarl, Paris, France
- German Research Centre for Artificial Intelligence, DFKI, Saarbrücken, Germany
- Information and Language Processing Systems, U. Amsterdam, Netherlands
- IZ Bonn, Germany
- Inst. For Information technology, Hyderabad, India
- LSI-UNED, Madrid, Spain
- Linguateca, Sintef, Oslo, Norway
- Linguistic Modelling Lab., Bulgarian Acad Sci
- NIST, USA
- Biomedial Informatics, Oregon Health and Science University, USA
- Research Computing Center of Moscow State U.
- Research Institute for Linguistics, Hungarian Academy of Sciences
- School of Computer Science and Mathematics, Victoria U., Australia
- School of Computing, DCU, Ireland
- UC Data Archive and School of Information Management and Systems, UC Berkeley, USA
- University "Alexandru Ioan Cuza", IASI, Romania
- U. Hospitals and U.of Geneva, Switzerland

# CLEF Steering Committee


- Maristella Agosti, University of Padova, Italy
- Martin Braschler, Zurich, Switzerland
- Amedeo Cappelli, ISTI-CNR & CELCT, Italy
- Hsin-Hsi Chen, National Taiwan U., Taipei, Taiwan
- Khalid Choukri, ELRA/ELDA, Paris, France
- Paul Clough, University of Sheffield, UK
- Thomas Deselaers, RWTH Aachen University, Germany
- 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, NIST, USA
- Gareth Jones, Dublin City University, Ireland
- Franciska de Jong, University of Twente, Netherlands
- Noriko Kando, NII, Tokyo, Japan
- Jussi Karlgren, SICS, 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
- Diana Santos, Linguateca, Sintef, Oslo, Norway
- 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, NIST, USA
- Christa Womser-Hacker, University of Hildesheim, Germany

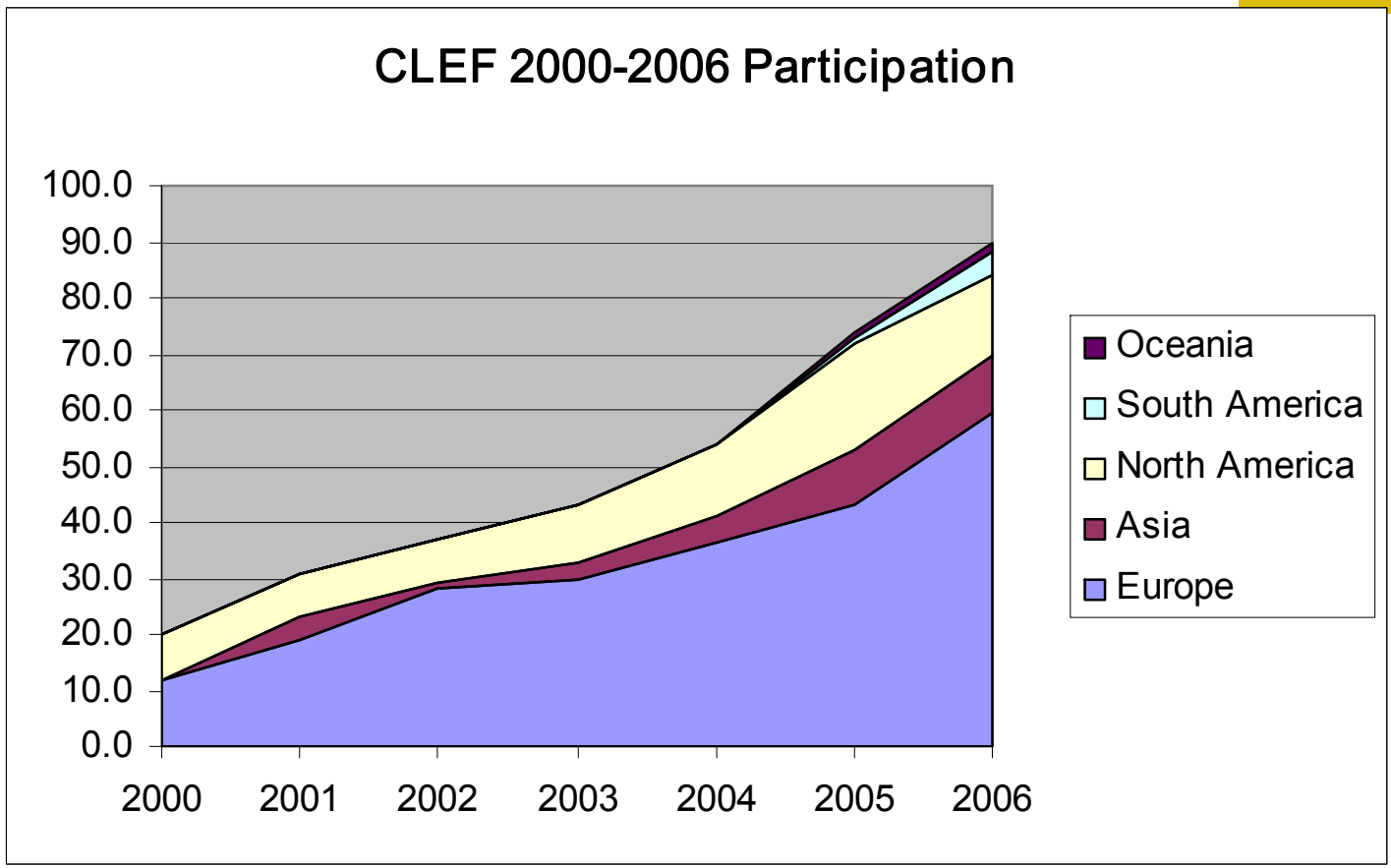
# CLEF 2006: Track Coordinators


- Ad Hoc: Giorgio Di Nunzio, Nicola Ferro and Thomas Mandl
- Domain-Specific: Maximilian Stempfhuber, Stefan Baerisch and Natalia Loukachevitch
- iCLEF: Julio Gonzalo, Paul Clough and Jussi Karlgren
- QA@CLEF: Bernardo Magnini, Danilo Giampiccolo, Fernado Llopis, Elisa Noguera, Anselmo Peñas and Maarten de Rijke
- ImageCLEF: Paul Clough, Henning Müller, Thomas Deselaers , Michael Grubinger, Thomas Lehmann, Allan Hanbury, and William Hersh
- CL-SR: Douglas W. Oard & Gareth J. F. Jones
- Web-CLEF: Krisztian Balog, Leif Azzopardi, Jaap Kamps, Maarten de Rijke
- GeoCLEF: Fredric Gey, Ray Larson, Mark Sanderson,

# CLEF 2006: Participating Groups


- Budapest U. Tech.&Economics, HU*
- Bulgarian Acad.Sci –TreeBank**
- California State U. SanMarcos, USA*
- CEA-LIST / LIC2M, France ***
- CELI, Italy
- Daedalus & Madrid Univs, Spain ***
- DFKI-Artificial Intelligence, DE***
- Dokuz Eylul U.,Turkey
- Dublin City U. - Comp.Sci., Ireland **
- ENSM - St Etienne, France*
- Hummingbird, Canada *****
- INSA Rouen, FR
- Inst.Infocomm Research, Singapore *
- IPAL-CNRS (IR2), Singapore ***
- **ITC-irst Trento, Italy ******
- Ist.Nac.Astrofisica, Optica, Electronica, Mexico*
- Imperial College, London, UK
- Indian Statistical Inst., India
- **Johns Hopkins U., USA ******
- JRC-ISPRA
- Lab Informatique Avignon, France
- Language Computer Corp., USA
- Language Tech. Research Centre, India
- LexiCLONE Inc.
- LIMSI-CNRS, France ***
- Linguateca-Sintef, Norway **
- Microsoft Asia
- Nat.Chiao-Tung U.-CS, Taiwan **
- Nat. Inst.Informatics, Japan **
- Nat.Taiwan U. - Comp-Sci, *****
- Oregon Health & Sci. U., USA **
- Priberam Informatica, Portugal *
- Queen Mary U. London, UK
- RWTH Aaachen-CS., Germany **
- RWTH Aachen - Med.Inf., DE**
- R2D2, Spain
- SUNY Buffalo – Informat, USA ***
- SICS, Sweden *****
- SYNAPSE Développement, France*
- Tech U. Chemnitz, Germany
- Tokyo Inst. Technology, Japan
- U. Hospitals Geneva, Switzerland **
- U.Alicante - Comp.Sci, Spain *****
- U.AI.I Cuza Iasi, Romania
- U.Amsterdam - Informatics, N *****
- U.Autonomous Puebla - CS, Mexico*
- U.Catolica Rio Grande do Sul, Brazil
- U. Computense Madrid, Spain
- U.Concordia - Comp.Sci, Canada*
- U.Coruna & U.Sunderland, ES/UK
- U. Essex & U.West Bohemia, UK/CZ
- U.Fed Sao Carlos, Brazil
- U.Freiburg – Pattern Recog., Germany
- U.Freiburg – Med.Inf., Germany
- U. & Hospitals Geneva, CH **
- U.Groningen - Inf.Sci, Netherlands*
- U.Hagen – IICS, Germany ***
- U.Hildesheim - Inf.Sci, Germany ***
- U.Indonesia - Comp.Sci, Indonesia *
- U.Jaen - Intell.Systems, Spain *****
- U.Liege - Elect.Eng.&CS, Belgium*
- U.Limerick - Comp. Sci, Ireland ***
- U.Lisbon – Informatics, Portugal **
- **U.Maryland - Comp.Sci, USA ******
- U.Melbourne – NICTA, Australia*
- U.Milan-Bicocca & U.Rome-Tor Vergata
- U. Nantes – Informatique, France*
- U.Neuchatel – Informatique, Switzerland *****
- U.Ottawa - IT & Eng, Canada*
- U.Politecnica Catalunya – TALP, Spain*
- U.Politecnica Valencia - Comp.Sci, Spain*
- U. Porto, Portugal
- U.Roma La Sapienza*
- U.Salamanca – REINA, Spain ****
- U.Sao Paulo, Brazil
- U.Sao Paulo & U.Fed Rio Grande do Sul, Brazil
- **U.Sheffield - Inf.Studies, UK ******
- U.Stockholm, NLP, Sweden **
- U.Stuttgart, Germany
- U.Texas at Dallas, USA
- U. Toulouse/CNRS, France
- U.Twente, The Netherlands ***
- U.Twente & U.Edinburgh, NL/UK
- U.West Bohemia, Czech Rep.
- U.Wolverhampton
- **UC Berkeley - IM&S-1, USA ******
- UNED-LSI, Spain *****
- U.New South Wales, Australia
- Vanguard Engineering, Mexico
- Wroclaw U. Technology, Poland

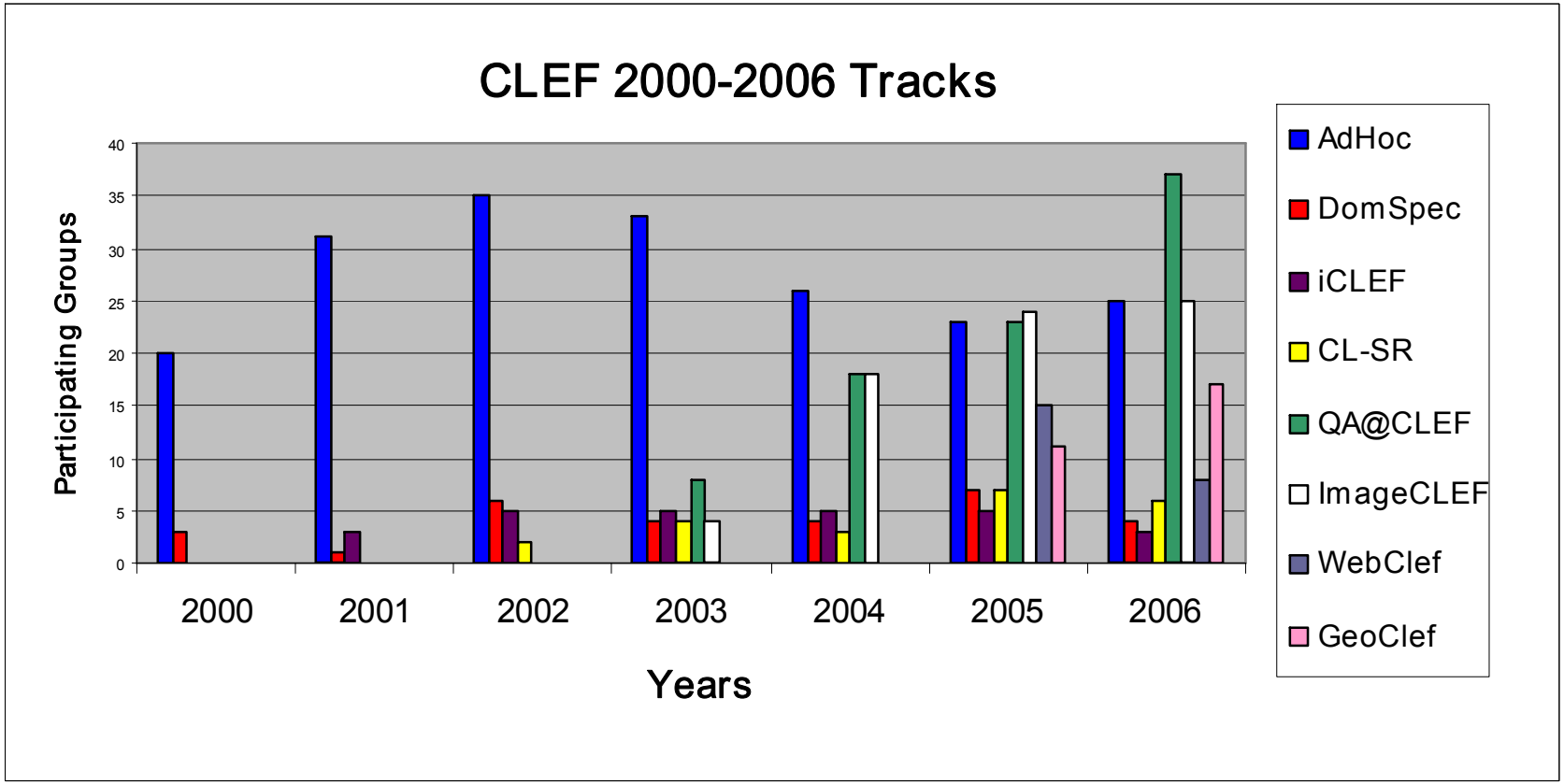
# CLEF: Growth in Participation



# No. of Participants per Track


- Ad Hoc:  25
- Domain-Specific - 4
- iCLEF – 3
- CL-SR - 6
- QA@CLEF - 37
- ImageCLEF - 25
- WebCLEF - 8
- GeoCLEF - 17

# CLEF 2000 – 2006 Tracks



# CLEF 2006 Document Collections


## Ad Hoc, QA@CLEF, iCLEF, GeoCLEF

- CLEF multilingual comparable corpus of more than 2M news docs in 12 languages: DE,EN,ES,FI,FR,IT,NL,RU,SV, PT, **BG** and **HU** (new in 2005)

## Domain-Specific

- The GIRT-4 social science database in EN and DE: more that 300,000 docs
- **The Russian Social Science Corpus:** almost 100,000 docs

## ImageCLEF

- St Andrews historical photographic archive: 28,000 images
- CasImage radiological medical database with case notes in FR and EN: 9,000
- **PEIR** 33,000 images, **MIR** 2,000, **PathoPic** 9,000
- **IRMA** collection in EN and DE for automatic medical image annotation: 10,000

## CL-SR

- **Malach collection** of spontaneous conversational speech derived from the Shoah archives: 589 hours

## WebCLEF

- **EuroGOV**, a multilingual collection of more than 2M webpages crawled from European governmental sites

# CLEF 2006 Topics


- **Ad hoc**
  - Mono- and Bi-: 50 topics in 13 languages
  - Multilanguage: 60 topics from CLEF 2003
- **Domain Specific**
  - 25 topics in 25 in EN, DE and RU
- **QA@CLEF**
  - 200 questions in 10 languages
- **ImageCLEF**
  - Ad Hoc 28 topics in 7 languages (All Fields) and 25 languages (title only)
  - Medical 25 topics: visual, text and visual, semantic; text in 3 languages
- **CL-SR**
  - x training topics and 25 eval. Topics in EN, CZ, FR, DE, ES
- **WebCLEF**
  - > 500 topics in 11 languages
- **GeoCLEF**
  - 25 topics in DE, EN, ES, PT

# CLEF 2006: Results


- Participation is up: 74 groups in 2005 (54 in 2004)
- Expansion of test-suites
- Great success of QA@CLEF and ImageCLEF
- Much interest in CL-SR, GeoCLEF and WebCLEF
- CLEF research community: synergy of diverse expertise partly consequence of new tracks – IR, NLP, Image Processing, Speech Processing, GIS, …
- CLEF 2005 Workshop 21-23 September, in conjunction with ECDL2005, >110 participants (ca 95 in 2004)

# CLEF Results in 10 Yrs


- Creation of strong CLIR research community (increase in participation over years )
- Strong profile (we are “known”)
- Promotion of research in key areas (multilingual IR; results merging; cross-language access in multimedia; interactive query formulation and results presentation)
- Encouraged take-up of techniques/resources between research groups
- Stimulated synergy between researchers from different areas (IR, NLP, Image Processing, User Interfaces, …)
- Literature: Working Notes, Proceedings and other publications report state-of-the-art plus emerging trends
- Production of language resources; test-suites

# CLEF in 2006: Ten Years Activity


## Focus on text retrieval

- monolingual/bilingual/multilingual document retrieval tasks
- mono- and cross-language IR on domain-specific data

## Focus on multi and mixed media retrieval

- mono-, bi- and multilingual text retrieval (Ad-hoc)
- scientific document retrieval (Domain-specific)
- interactive cross-language retrieval (iCLEF)
- multiple lang. question answering (QA@CLEF)
- cross-lang. retrieval on image collections (ImageCLEF)
- cross-lang. speech retrieval (CL-SR)
- multilingual web retrieval (WebCLEF)
- cross-lang. geographic retrival (Geo CLEF)

# CLEF 2005 Proceedings


**Accessing Multilingual Information Repositories**
6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers
Series: Lecture Notes in Computer Science ,
Vol. 4022
Sublibrary: Information Systems and Applications, incl. Internet/Web, and HCI
Peters, C.; Gey, F.; Gonzalo, J.; Mueller, H.; Jones, G.; Kluck, M.; Magnini, B.; de Rijke, M. (Eds.)
2006, XXI, 1013 p., Softcover
ISBN: 3-540-45697-X

# CLEF Objectives


- Stimulate the development of multilingual IR systems for European languages
- To create a CLIR community
- Construct publicly available test-suites
- Conducting annual evaluation campaigns
- Designing tracks/tasks to meet emerging needs and to stimulate research in the”right” direction

# CLEF in 2002: Six Years Activity


## Focus on text retrieval

- monolingual/bilingual/multilingual document retrieval tasks
- mono- and cross-language IR on domain-specific data

## Growth in participation

- 13 groups in 1997 – ca 40 groups in 2002
  - more European groups – more industrial groups
- annual workshops

## Creation of test collection

- comparable corpus in 8 languages; queries in 12
- scientific texts collection in German and French
- data and relevance assessments from past campaigns are available to registered participants free-of-charge

# What the User wants (aot)


- Larger test collection (more languages and more data)
- Different text types (e.g. structured data)
- More task variety (question-answering, web-style queries, text categorization)
- Ways to test retrieval with multimedia data
- More focus on user satisfaction issues (e.g. query formulation, results presentation)

# CLEF in 2006: Growth in participation


- 13 groups in 1997 – ca 40 groups in 2002
  - more European groups – more industrial groups
- More than 90 groups in 2006 (110 registered)
  - from (almost) all continents – few industrial groups

# CLEF in 2006: Creation of test collection


## 2002

- comparable corpus in 8 languages; queries in 12
- scientific text collection in German and French
- data and relevance assessments from past campaigns are available to registered participants free-of-charge

## 2006

- CLEF multilingual comparable corpus of more than 2M news docs in 12 languages: DE,EN,ES,FI,FR,IT,NL,RU,SV,PT,BG and HU
- GIRT-4 social science database in EN and DE: more that 300,000 docs; 2 Russian Social Science Corpora: 250,000 docs
- IAPR photo collection, captions in EN & DE; LTU-Tech images for non-medical annotation
- CasImage radiological medical database with case notes in FR and EN: 9,000; PEIR 33,000 images, MIR 2,000 images, PathoPic 9,000 images; IRMA collection in EN and DE for automatic medical image annotation:10,000 images
- Malach collection of conversational speech derived from the Shoah archives EN & CZ (speech recognition, controlled vocab. Descriptors, word lattices)
- EuroGOV, a multilingual collection of more than 2M webpages crawled from European governmental sites

# CLEF:Overall Results


- Stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- Study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- Documented improvement in system performance for cross-language text retrieval systems
- Creation of a large set of empirical data about multilingual information access from the user perspective
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- Creation of important, reusable test collections for system benchmarking
- Building of a strong, multidisciplinary research community

# CLEF in 2006
# What haven't we done ?


- Where are the systems?
- We've forgotten the users
- (Are there any users?)

# What the User wants (aot)


- Larger test collection (more languages and more data)
- Different text types (e.g. structured data)
- More task variety (question-answering, web-style queries, text categorization)
- Ways to test retrieval with multimedia data
- More focus on user satisfaction issues (e.g. query formulation, results presentation)

# Points for Discussion


- What new tasks/evaluation methodologies are needed to address more advanced information requirements?
- How can we best reduce the gap between research and application communities?

- What are we doing wrong?
- What should we be doing?
- Who are the users?
  - Is there a use case?

# The Future of CLEF


???

2003

Can we survive?!

# The Future of CLEF


???

CLEF 2004

It's looking fine!

# The Future of CLEF


???

CLEF 2005

Are we doing too much?!

# The Future of CLEF


???

CLEF 2006

## Is 2007 the end, my friend?