



PROMISE

Participative Research labOratory for Multimedia and
Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 6.1

Report on the outcomes of the first year evaluation activities

Version 1.1, 31 August 2011



Document Information

Deliverable number:	D6.1
Deliverable title:	Report on the outcomes of the first year evaluation activities
Delivery date:	31/08/2011
Lead contractor for this deliverable	HES-SO
Author(s):	Theodora Tsikrika, Henning Müller, Pamela Forner, Madlen Friesseke, Florina Piroi, Maristella Agosti, Emanuele Di Buccio, Richard Berendsen
Participant(s):	All
Workpackage:	WP6
Workpackage title:	Evaluation activities
Workpackage leader:	HES-SO
Dissemination Level:	PU – Public
Version:	1.1
Keywords:	Evaluation activities, CLEF conference, CLEF Labs, CLEF-IP, ImageCLEF, medical retrieval, Cultural Heritage workshop, CLEF campaign website quality analysis, ImageCLEF scholarly impact

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	30/05/11	Draft	HES-SO	First draft
0.2	12/07/11	Draft	All partners	Sent for internal review
0.3	04/08/11	Draft	HES-SO	Changes after internal review
0.4	05/08/11	Draft	HES-SO	Sent for internal review
1.0	12/08/11	Draft	HES-SO	Revised after internal review
1.1	31/08/11	Final	HES-SO	Final version

Abstract

This deliverable reports on the outcomes of the evaluation activities in the first year of PROMISE. PROMISE organizes experimental evaluation activities for multilingual and multimedia information access systems at an international level and on an annual basis; these activities are embedded in the Cross Language Evaluation Forum (CLEF), a renowned evaluation framework. As of 2010, CLEF consists of an annual conference on experimental evaluation and a series of participative benchmarking activities referred to as labs. Therefore, this report presents the outcomes of the CLEF conference and labs, with particular focus on the CLEF labs organized for the three domains of the PROMISE use cases, i.e., cultural heritage, intellectual property, and multimedia (mainly image and text) medical retrieval. We discuss the lessons learned so as to monitor the evolution of these evaluation activities and intercept emerging trends with the goal to establish a point of reference for future evaluation campaigns based on measurable criteria, deliver solutions to the encountered problems, and advance the defined use cases. In addition to the experimental evaluation activities, we also report on the activities performed for the evaluation of the quality and impact of information and knowledge resources generated by CLEF. The deliverable concludes with an outlook on the evaluation activities for the second year of PROMISE.

Table of Contents

Document Information	2
Abstract	3
Table of Contents	4
Executive Summary	6
1 Introduction.....	9
2 Overview of the first year evaluation activities.....	11
2.1 CLEF 2010 Conference and Labs	12
2.1.1 CLEF 2010 Conference	12
2.1.2 CLEF 2010 Labs	13
2.1.3 Participation to the CLEF 2010 Labs.....	15
2.2 Main advancements	15
2.3 Main trends	16
2.4 Main problems from an organizational point of view	16
3 Outcomes of evaluation activities: CLEF 2010 labs test collections	17
3.1 Collections	17
3.2 Topics.....	18
3.3 Ground truth.....	18
4 Outcomes of the evaluation activities for the “Visual Clinical Decision Support” Use Case 20	
4.1 Medical Modality Detection Task.....	20
4.2 Medical Image Retrieval Task	22
4.3 Medical Case Retrieval Task.....	23
4.4 Summary of the outcomes of the “Visual Clinical Decision Support” Use Case	25
5 Outcomes of the evaluation activities for the “Search for Innovation” Use Case	27
5.1 Prior Art Candidates Search Task.....	28
5.2 Patent Classification Task.....	29
5.3 Other activities for the “Search for Innovation” use case	30
5.4 Summary of the outcomes of the “Search for Innovation” use case.....	31
6 Outcomes of the evaluation activities for the “Unlocking Culture” Use Case	33
6.1 CHiC2011.....	33
6.2 Reports: State-of-the-Art Evaluation of Digital Libraries in the Cultural Heritage domain	34
6.2.1 Report on Use Case Components in Cultural Heritage Information Systems ...	34

6.2.2	Report on Evaluation in Cultural Heritage	35
7	Further outcomes of evaluation activities: information and knowledge resources	36
7.1	CLEF campaign: scholarly impact analysis	37
7.1.1	ImageCLEF scholarly impact: bibliometric analysis method.....	37
7.1.2	ImageCLEF scholarly impact: results	38
7.1.3	Conclusions	41
7.2	CLEF campaign website: web design qualitative analysis	42
7.2.1	Website quality model and quality analysis methodology	43
7.2.2	CLEF campaign website qualitative analysis	45
8	Outlook on future evaluation activities: CLEF 2011	47
9	References.....	49
	Appendix I: Questionnaires sent to CLEF 2010 Labs organizers	52
	Appendix II: Participation in the CLEF 2010 labs	64
	Appendix III: Main outcomes of the CLEF 2010 Labs	67
	Appendix IV: CLEF 2010 Labs Test Collections.....	75
	Appendix V: Qualitative analysis of CLEF campaign website	85
	Appendix VI: CLEF 2010 website statistics.....	90
1.	CLEF 2010 website: March 2010 – December 2010 statistics.....	90
2.	CLEF 2010 website: January 2011 – June 2011 statistics.....	92

Executive Summary

This deliverable presents the main outcomes of the evaluation activities in the first year of PROMISE, i.e., the outcomes of (i) the experimental evaluation activities performed in the context of the CLEF conference and labs, with particular focus on the activities of the three PROMISE use cases, and (ii) the activities performed for the evaluation of the quality and impact of information and knowledge resources generated by the CLEF evaluation activities. The deliverable concludes with an outlook on the evaluation activities for the second year.

• Evaluation activities in CLEF 2010: Conference and Labs

PROMISE organizes experimental evaluation activities for multilingual and multimedia information access systems at an international level and on an annual basis; these activities are embedded in CLEF. As of 2010, CLEF consists of an annual conference on experimental evaluation and a series of participative benchmarking activities referred to as labs. We first present a short overview of the **CLEF 2010 conference** together with a short description of the **CLEF 2010 labs** and the participation to them. To gain insights on the outcomes of the CLEF 2010 labs and to form a point of reference for monitoring the evolution and progress of the CLEF labs over the coming years, we then present the results of two questionnaires sent to the CLEF 2010 lab organizers. These results can be summarized as follows:

1. **Tasks:** A total of 13 tasks were investigated in the CLEF 2010 labs: four classification tasks, four (ad-hoc) information retrieval tasks, whereas the rest encompass a wide variety of tasks, namely question answering, document filtering, document clustering and information extraction, expert search, and log analysis.
2. **Main advancements:** Overall, the observed tendencies in the evolution of tasks over the last two years are closely aligned with the PROMISE objectives towards larger data sets consisting of multimedia and multilingual content and more realistic tasks. (i) Seven out of the nine tasks that also ran in 2009 employed larger collections. (ii) In many cases, additional resources were provided. (iii) In most cases, the number of topics (or classes) was increased. (iv) Efforts were made towards making the tasks more realistic.
3. **Main trends in the participants' approaches:** (i) The use of external resources appears to be beneficial. (ii) Several combination of evidence approaches are applied so as to take into account these resources, but also to consider domain-specific evidence, as well as the evidence obtained from the multiple media and languages in these complex environments. (iii) Document classification can be used as a filtering step to enhance retrieval, while topic classification can be used in order to apply different approaches to different types of queries.
4. **Main problems:** (i) Ground truth creation: large amount of human effort; difficulties in recruiting volunteers when funding is not available; disagreement among assessors; and the quality of annotations when crowdsourcing is employed. (ii) Copyright management of the data in the collections. (iii) Low participation rate compared to the number of registrations. (iv) Lack of funding for performing in-depth analysis of the collected experimental data. PROMISE aims to address these issues through the automation of experimental evaluation, the curation, preservation, and enrichment of experimental data, the development of well-defined and compelling use cases, and the support of participants through the open evaluation infrastructure.

5. Test collections generated by the CLEF 2010 labs.

- a. **Collections:** The CLEF 2010 labs collections are evidence of the large size of the data sets already employed in the PROMISE evaluation activities. The continuous update of existing data sets manifests a tendency to increase the volume of data.
- b. **Topics:** Topic creation is an important step in the evaluation campaign cycle and is accompanied by significant challenges in not only creating topics that reflect realistic user information needs, but that these topics are also scientifically feasible and challenging at the same time. The development of well-defined use cases supports the topic creation process. The number of topics is crucial in ensuring the reliability of the experimental outcomes, but is ultimately determined by the effort required in creating the ground truth.
- c. **Ground truth:** Ground truth creation is one of the steps in the evaluation campaign that will benefit tremendously by the automation in the experimental evaluation process currently being investigated by PROMISE. Out of the 13 tasks, one did not require ground truth, four exploited existing annotations in their collections to automatically generate relevance assessments, whereas the remaining eight tasks employed human assessors. Four of these eight tasks used crowdsourcing, while the other four enlisted the help of 3-12 human assessors. The human effort required to generate these relevance assessments varies greatly based on the nature and difficulty of the task, but can reach up to 300 hours for a single task.

• Evaluation activities for PROMISE use cases

We then present the conclusions of and lessons learned from the evaluation activities for the three PROMISE use cases. Steps towards addressing the identified problems and providing suitable solutions, as well as efforts to capitalize on the gained experience and knowledge so as to improve these evaluation activities are taking place for next year's evaluation activities, including in some cases collaboration with other evaluation campaigns.

1. "Visual Clinical Decision Support" Use Case (Medical retrieval task at ImageCLEF lab)

- (a) *The task remains very popular.* Even in its seventh edition, it attracts a high number of registrations and participations. There was an increase in the number of submitted runs, 155 in total, the highest number of runs submitted in any of the CLEF labs.
- (b) *More research is necessary for the effective and robust combination of evidence from different modalities.* Multimodal approaches are the most effective for the modality detection and medical image retrieval tasks, whereas further research is needed for the medical case retrieval task.
- (c) *Interactive retrieval is still being used only by a very small number of participants, although it does have the potential to improve retrieval effectiveness.* For the medical case retrieval, the best results were obtained with a textual retrieval approach when using relevance feedback. To encourage research in this area, a medical user-oriented (interactive) image retrieval task is organized for CLEF 2011.
- (d) *Inter-rater agreement can be low for topics with few relevant images.* Given that the relative rankings of the groups were vastly unchanged when using the assessment of different judges aside from topics with low number of relevant images, efforts should be made to remove topics with very few relevant images.

2. “Search for Innovation” Use Case (CLEF-IP Lab)

- (a) *The CLEF-IP lab does not evaluate the full “search for innovation” process, as understood by the professional patent searchers.* This is normal and agreed given that CLEF-IP focuses on the cross-lingual full text retrieval evaluation; nevertheless, a better communication to professional users is needed to convey the objectives of the CLEF-IP lab. Furthermore, any “search for innovation” performed by professional patent searches is an iterative process, with continuous query refinement; this behaviour is not currently reflected in the CLEF-IP lab.
- (b) *The “Search For Innovation” Use Case is very domain specific.* Chemistry, for example, has a retrieval process on its own and cross-lingual search is useless in such a domain where there exists a universal language. The aim is to expand this use case, through our collaboration with NIST on the organization of TREC-CHEM, by bringing that evaluation campaign into PROMISE.
- (c) *Classification at IPC (sub)class level performs well across most participating groups, indicating that the problem may be trivial.* A task for IPC classification at sub-group level is introduced in 2011.
- (d) *Users find the patent search hard to understand and lack motivation to participate.* Users need to become more directly involved, e.g., by showing them how each system works. To this end, PROMISE supports the PatOlympics evaluation campaign.

3. “Unlocking Culture” Use Case (Cultural Heritage workshop at CLEF 2011: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage)

The evaluation activities for this use case centre on the forthcoming CLEF 2011 Cultural Heritage (CH) workshop. The aim of this workshop is to establish standard evaluation criteria and methods within this domain by: (i) establishing what makes searching in the CH domain distinct from other domains, (ii) gathering existing use cases for multilingual information access in the CH domain, (iii) reviewing existing evaluation resources studies within the CH domain, (iv) proposing appropriate methodologies for evaluating multilingual information access to CH resources, and (v) defining multiple concrete evaluation tasks modelled on IR evaluation initiatives such as CLEF, TREC or INEX.

• Evaluating information and knowledge resources generated by CLEF

In addition to the experimental evaluation activities, we also evaluated the quality and impact of two information and knowledge resources generated by CLEF: (i) the ImageCLEF publications and (ii) the CLEF campaign website. The first study indicates ImageCLEF's significant scholarly impact through the substantial numbers of its publications and their received citations. Our goal is to expand this preliminary analysis by including additional ImageCLEF publications and to also perform a similar analysis for the whole of CLEF. The second study indicates that the quality of the CLEF campaign website is on average perceived as good and its content as reliable. Characteristics relating to the management aspects of the website obtained a good value, mainly due to the high frequency of updates and continuous availability, but its quality can be further increased by improving the monitoring activities. Finally, while an expert user familiar with the workflow of an evaluation campaign can successfully reach the information he is looking for, the informative structure should be improved by making the relationships among informative resources explicit.

1 Introduction

PROMISE aims at advancing the experimental evaluation of complex multimedia and multilingual information systems through a virtual laboratory for conducting participative research and experimentation. To this end, PROMISE organizes regular experimental evaluation activities for multilingual and multimedia information systems at an international level and on an annual basis. These evaluation campaigns enable the reproducible and comparative evaluation of new approaches, algorithms, theories, and models, through the use of standardised resources and common evaluation methodologies within regular and systematic evaluation cycles. Such organised benchmarking activities have been widely credited with contributing tremendously to the advancement of information access and retrieval by providing access to infrastructure and evaluation resources that support researchers in the development of new approaches, and encouraging collaboration and interaction between researchers both from academia and industry.

Evaluation campaigns are predominantly based on the Cranfield paradigm [Cleverdon, 1959] of experimentally assessing the worth and validity of new ideas in a laboratory setting through the use of *test collections*, each consisting of (i) a *collection* of documents, (ii) a set of user requests (*topics*), and (iii) a set of relevance judgements (*ground truth*). Our goal is to advance this traditional way of conducting evaluation campaigns by relying on large data sets, tackling realistic use cases and evaluation tasks designed for compelling user and industrial needs, advancing and automating the evaluation process to better support the envisioned tasks and use cases, providing a proper evaluation infrastructure, producing information and knowledge resources from the collected experimental data, and involving Europe-wide large researcher and developer communities with multidisciplinary competencies.

This deliverable reports on the outcomes of the concrete experimental evaluation activities that have taken place during the first year of PROMISE, with particular focus on the evaluation campaigns organized for the three domains of the PROMISE use cases, i.e., cultural heritage, intellectual property, and multimedia (mainly image and text) medical retrieval. These evaluation campaigns are conducted under the auspices of the Cross Language Evaluation Forum (CLEF), a renowned evaluation framework. As of 2010, CLEF consists of an annual conference on experimental evaluation and a series of participative benchmarking activities referred to as labs. Therefore, this report presents the outcomes of the CLEF conference and labs, in particular those labs based on the PROMISE use cases, and discusses the lessons learned so as to (i) monitor the evolution of these evaluation activities and intercept emerging trends with the goal to establish of point of reference for future evaluation campaigns based on measurable criteria, (ii) deliver solutions to the encountered problems, and (iii) advance the defined use cases.

Besides fostering, supporting, and coordinating experimental evaluation activities, PROMISE also aims to curate, preserve, and enrich the information and knowledge resources resulting from such activities, such as experimental data, methodologies, and publications, as well as the possible relationships among them; the ultimate goal is to provide access to such resources so that they can be put to use towards the research and development of multimedia and multilingual information access systems. In this first year of

PROMISE, we have conducted two separate evaluations on the quality and impact of such resources: (i) the CLEF-derived publications and (ii) the Cross Language Evaluation Forum (CLEF) websites.

This deliverable is structured as follows. Section 2 provides an overview of the first year evaluation activities by discussing the main outcomes of and the lessons learned from the CLEF 2010 conference and labs. Section 3 focuses on one of the main outcomes of these experimental evaluation activities, the test collections generated by the CLEF 2010 labs. Sections 4, 5, and 6 provide a more detailed analysis of the outcomes of the evaluation activities for the three PROMISE Use Cases, respectively. Section 7 presents the outcomes of the evaluation of the quality and impact of the information and knowledge resources generated as a result of the experimental evaluation activities. Section 8 concludes by providing an outlook on the current status of the CLEF 2011 conference and labs.

2 Overview of the first year evaluation activities

PROMISE organizes experimental evaluation activities for multilingual and multimedia information access systems at an international level and on an annual basis; these activities are embedded in CLEF, the Cross Language Evaluation Forum initiative. As of 2010, CLEF consists of an annual conference on experimental evaluation and a series of participative benchmarking activities referred to as labs. The CLEF Conference on Multilingual and Multimodal Information Access Evaluation takes place in September of each year, while the evaluation activities of the CLEF Labs run on an annual cycle during the twelve month period preceding the conference and culminate in workshops taking place in conjunction with the conference.

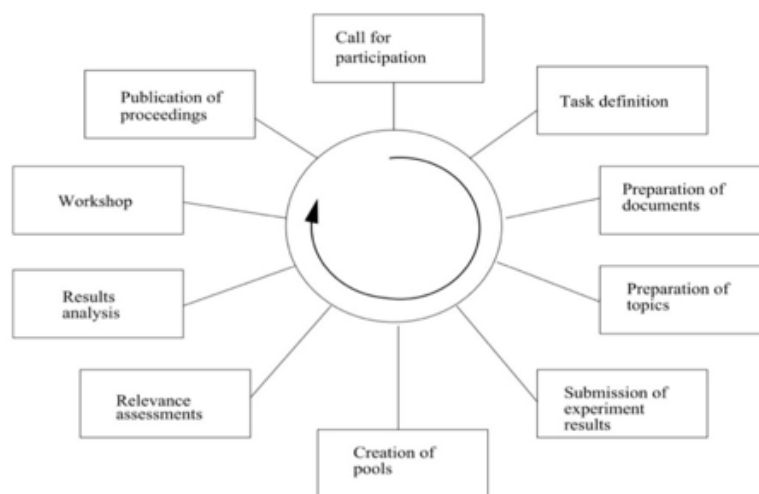


Figure 1: Annual cycle of activities in an evaluation campaign [adapted from <http://trec.nist.gov/presentations/TREC2004/04intro.pdf>]

The typical annual evaluation cycle for CLEF Labs is depicted in Figure 1. It begins with a call for participation followed by an expression of interest from research groups and their registration. Each lab may consist of one or more evaluation tasks defined by the lab organisers, who are also responsible for preparing the collections and topics and providing them to the participants. The participants use these datasets to run their experiments and produce system outputs in standard format (called *runs*), which are then submitted to the lab so as to be evaluated. Their evaluation is based on relevance assessments for each topic performed either for the whole collection, or more commonly for a subset of the collection corresponding to pools of documents from the submitted runs. Evaluation measures are used for assessing the runs' performance based on the number of relevant documents found. Results are released and analysed prior to the workshop so as to share insights and discuss findings. Finally, the activities and results are published in the labs' working notes and/or workshop proceedings.

This section provides an overview of these evaluation activities during the first year of PROMISE: Section 2.1 briefly reports on the outcomes of the CLEF 2010 conference and

labs. To gain insights on the outcomes of these evaluation activities and in particular of the CLEF 2010 labs, a questionnaire (see Appendix I: Questionnaires sent to CLEF 2010 Labs organizers) was prepared and sent to the CLEF 2010 lab organizers in May 2011. Some of the results of this questionnaire are presented in the following sections (Sections 2.2-2.4) and will form a point of reference for monitoring the evolution and progress of the CLEF labs over the coming years. By tracking the changes and intercepting the trends emerging in these labs, PROMISE will be able to react more effectively and deliver appropriate solutions that advance experimental evaluation by moving it from an handicraft process to a mostly automatic one. Further information on the CLEF 2010 labs can be found in the electronic notebook papers which are available online at the CLEF 2010 website (<http://www.clef2010.org/>).

2.1 CLEF 2010 Conference and Labs

The CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation – the first event organized by PROMISE – represented an innovation of the “classic CLEF” format and an experiment aimed at understanding how next generation evaluation campaigns might be structured. The main concern was how to innovate CLEF while still preserving its traditional core business, namely the benchmarking activities carried out in the various tracks and tasks. The major novelty was that CLEF was made an independent four-day event, i.e., it was no longer organized in conjunction with the European Conference on Research and Advanced Technology for Digital Libraries (ECDL), where CLEF has run as a two and half day workshop for the previous ten years. CLEF 2010 thus consisted of two main parts: (i) a *peer-reviewed conference* on experimental evaluation, which innovated the CLEF tradition and aimed at advancing the evaluation of complex multilingual and multimodal information systems in order to support individuals, organizations, and communities who design, develop, employ, and improve such systems; and (ii) a *series of labs*, which continued the CLEF tradition of community-based evaluation.

Deliverable 7.2 “First PROMISE Annual Conference and Proceedings” provides a detailed overview of the CLEF 2010 conference, a description of the labs, and a report on the participation to these events. Below, we summarize the main outcomes of the CLEF 2010 conference and labs and also analyze the participation to the CLEF 2010 labs.

2.1.1 CLEF 2010 Conference

The CLEF 2010 conference on Multilingual and Multimodal Information Access Evaluation – the first event organized by PROMISE - held at the University of Padua from 20th to 23rd September. In summary:

1. The CLEF 2010 conference represented an innovation of the traditional structure of evaluation campaigns with a renewed organizational structure.
2. There were 21 submissions (17 full papers and 4 short papers) and each paper had, on average, 4 reviews. In total, 12 papers were accepted (8 full papers and 4 short papers) with an overall acceptance rate of 57%. The papers accepted for the conference comprised research on resources, tools, and methods, experimental collections and data sets, and evaluation methodologies and metrics.
3. There were two keynote talks and two panels.

4. All accepted papers, invited talks, and panels were published by Springer in their Lecture Notes for Computer Science series (volume 6360); these proceedings were distributed to all the attendees at the time of the conference.
5. This first PROMISE event attracted a large number of participants, approximately 140 researchers, with most of them staying for the full four days, indicating the success of this innovation of the “CLEF format”.

2.1.2 CLEF 2010 Labs

The CLEF 2010 Labs continued the CLEF tradition of community-based benchmarking and complemented it with workshops on emerging issues in evaluation methodology. In fact, two different forms of labs were offered: labs could either be run as *benchmarking activities* “campaign-style” during the twelve month period preceding the conference, or as *exploration workshops* by adopting a more “workshop-style” format that could explore issues of information access evaluation and related fields. There were 9 lab proposals: 5 were accepted as “campaign-style” or benchmarking activities and 2 were accepted as exploration workshops, resulting in an acceptance rate of 7/9 (=77%).

The five benchmarking evaluations that ran as labs in CLEF 2010 were:

1. **CLEF-IP**: a benchmarking activity on intellectual property [Piroi, 2010c]. CLEF-IP, sponsored by the Information Retrieval Facility (IRF) in Vienna, was a follow-up to a CLEF 2009 track. There were two tasks in 2010: the *Prior Art Candidates Search Task* [Piroi, 2010a] to find patent documents that are likely to constitute prior art to a given patent application, and the *classification task* [Piroi, 2010b] which aimed to classify a given patent document according to the IPC codes.
2. **ImageCLEF**: a benchmarking activity on image retrieval and annotation. It was the eighth running of this track in CLEF. There were four tasks in 2010: *Medical Retrieval* [Müller et al., 2010] from 77,000 images from articles published in Radiology and Radiographics, *Photo Annotation* [Nowak & Huiskes, 2010] of a MIR Flickr 25,000 database of consumer photos with multiple annotations, *Robot Vision Challenge* [Pronobis et al., 2010], and *Wikipedia Image Retrieval* [Popescu et al., 2010] using 237,000 Wikipedia images that cover diverse topics of interest and are associated with unstructured and noisy textual annotations in English, French, and German.
3. **PAN**: a benchmarking activity on uncovering plagiarism, authorship, and social software misuse. PAN ran at CLEF for the first time, following three previous workshops at other conferences. It was sponsored by Yahoo! Research and had two tasks, namely the *detection of plagiarism* [Potthast et al., 2010a] and the *detection of Wikipedia vandalism* [Potthast et al., 2010b].
4. **ResPubliQA (QA@CLEF)**: a benchmarking activity on question answering using multilingual political data [Peñas et al., 2010]. This was the eighth year for multilingual question answering in CLEF. Similar to the ResPubliQA version in CLEF 2009, the lab used the Europarl Corpus, and had seven monolingual tasks for English, French, German, Italian, Portuguese, Spanish and Romanian.
5. **WePS**: a benchmarking activity on Web People Search. WePS focused on person name ambiguity and person attribute extraction on Web pages [Artiles et al., 2010] and on Online Reputation Management (ORM) for organizations [Amigó et al., 2010],

again dealing with the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. This was the lab's first year at CLEF, following two previous workshops at other conferences.

The two exploration workshops that ran as labs in CLEF 2010 were:

6. **CriES** addressed the problem of multi-lingual expert search in social media environments [Sorg et al., 2010]. The main topics were multi-lingual expert retrieval methods, social media analysis with respect to expert search, selection of data sets and evaluation of expert search results. Papers reporting on experiments or proposals for possible benchmarking activities were invited.
7. **LogCLEF** aimed at exploring methodologies for studying search engine log files [Mandl et al., 2010]. To this end, it investigated the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. The different log sets were used, The European Library (TEL) logs, and the Deutscher Bildungsserver (DBS) logs, a quality controlled Internet directory for educational resources. Participants were invited to investigate a variety of questions with the end goal of defining a benchmarking task for follow-on labs.

The results of the experiments conducted within CLEF 2010 labs were presented and discussed as sessions of half a day, one full day or two days at the CLEF 2010, 22-23 September, Padua, Italy. These sessions were run in parallel covering two days, while a general poster session was arranged at the end of the second day, where all participants from all the different Labs had the opportunity to present their work. These sessions play an important role by providing the opportunity to all the groups that participated in the labs to get together to compare approaches and exchange ideas.

The CLEF 2010 Labs consist of a total of 13 tasks listed in Table 1 below: four of them are Classification tasks, four of them are (ad-hoc) Information Retrieval tasks, whereas the rest encompass a wide variety of tasks such as Question Answering, Document Filtering, Document Clustering and Information Extraction, Expert Search, and Log Analysis.

Table 1: CLEF 2010 Labs and their tasks.

Lab	Task(s)
CLEF-IP	Patent Classification
	Prior Art Candidates Search
ImageCLEF	Medical image retrieval
	Photo Annotation
	Robot Vision
	Wikipedia image retrieval
PAN	Plagiarism Detection
	Wikipedia Vandalism Detection
ResPubliQA	Paragraph Selection (PS), Answer Selection (AS)
WePS	Online reputation management
	Clustering, Attribute Extraction
CriES	CriES Pilot Challenge

LogCLEF	LogCLEF
---------	---------

2.1.3 Participation to the CLEF 2010 Labs

The CLEF 2010 evaluation activities have achieved high visibility. Out of the 200 research groups initially registered, about 110 (mostly from Europe) submitted runs. Participation per lab: CLEF-IP 19, ImageCLEF 49, PAN 31, MLQA 24, WePS 21, CriES 7, and LogCLEF 19. Table 8 in Appendix II: Participation in the CLEF 2010 labs provides a more detailed breakdown on the number of registrations, participations, and return participations per task.

The most established tasks, i.e., those running for a number of years either under the auspices of CLEF or under other initiatives, attracted the most registrations. The most popular CLEF 2010 Lab was ImageCLEF able to attract many participants not only from Europe but also from the United States and other countries. The participation rate (i.e., the number of registered research groups that actually submitted their results to the lab) is on average 48%, with the lowest for the Robot Vision task at ImageCLEF (7 participants out of 43 registrations), and the highest of 100% for Cries and WePS-ORM. Return participations from the previous year are on average around 40%, indicating that a large number of researchers rely year after year on the resources created in the context of the CLEF evaluation activities.

The number of submissions varies greatly per task, with most having a couple of dozen submitted runs, with the exception of two of the ImageCLEF tasks that manage to attract over 120 submissions. These numbers of submitted experiments indicate the scale of experiments that the PROMISE evaluation infrastructure should handle. Furthermore, the different submission systems employed by each lab is further evidence to the necessity of a unified evaluation environment and infrastructure currently developed in PROMISE.

2.2 Main advancements

Given that for the majority of tasks in the CLEF 2010 labs (9 out of 13 tasks), this was not the first time they ran, it is worth noting the main differences and advancements compared to 2009. This comparison enables us to monitor both the progress made, as well as the main tendencies in the evolution of such evaluation activities. Table 9 in Appendix III: Main outcomes of the CLEF 2010 Labs presents the main differences between the two years as pointed out by the task organizers.

Most of the tasks (7 out of the 9 tasks that also ran in 2009) employed larger collections, either by updating existing collections or creating new ones from scratch. In many cases, additional resources were provided (e.g., Flickr user tags in the Photo Annotation task and Wikipedia articles in the Wikipedia image retrieval task). This also introduced a shift in the research objectives investigated in these tasks towards combination of evidence approaches that exploit these resources. In most cases, the number of topics (or classes) was increased, with the exception of the CLEF-IP retrieval task, where fewer topics compared to the previous year were considered. Efforts were also made towards making the tasks more realistic, e.g., by creating topics that correspond more closely to real practice (CLEF-IP retrieval and case-based medical image retrieval) and by defining the system output in a way that aims to fulfil more realistic user requirements (ResPubliQA and the Clustering and Attribute Extraction task at WePS).

Overall, the observed tendencies in the evolution of tasks over the two years are closely aligned with the PROMISE objectives towards larger datasets consisting of multimedia and multilingual content and more realistic tasks.

2.3 Main trends

Table 10 in Appendix III: Main outcomes of the CLEF 2010 Labs presents the main trends in the approaches employed by the participants, as well as the main outcomes of their experiments. Given the high heterogeneity of the tasks, the main purpose of this analysis is to not to identify trends and tendencies across tasks, but to establish a point of reference for monitoring the trends within each task over the coming years. Nevertheless, there are some overall tendencies that can be mentioned. First of all, the use of external resources appears to be beneficial; such resources include general-purpose ontologies (such as WordNet and DBpedia) or domain-specific ones (e.g., medical ones), Web search results related to the current query, and Flickr images and tags related to visual queries. The challenge of course is to integrate these resources to the applied retrieval models. To this end, several different combination of evidence approaches are applied so as to take into account these resources, but also to consider domain-specific evidence, such as patent-metadata in the case of CLEF-IP and the social graph in the case of CriES, as well as the evidence obtained from the multiple media and languages in these complex environments. Finally, document classification can be used as a filtering step to enhance retrieval, while topic classification can be used in order to apply different approaches to different types of queries.

2.4 Main problems from an organizational point of view

Most of the problems identified by task organizers concern the ground truth creation and include the large amount of human effort required and which is often underestimated when planning the task, the difficulties in recruiting volunteers when funding is not available, the disagreement among assessors when multiple judges are used per topic, and finally the quality of annotations when crowdsourcing is employed. Additional problems include the copyright management of the data in the collections, the low participation rate compared to the number of registrations, and the lack of funding for performing in-depth analysis of the collected experimental data.

PROMISE aims to address the issue of ground truth generation through the automation of experimental evaluation and the issue of the analysis of experimental results through the curation, preservation, and enrichment of experimental data. PROMISE can also contribute towards the increase of the rate of participation by promoting evaluation tasks that correspond to well-defined and compelling use cases, and thus stimulate research and development in the related fields, and also by supporting participants through the open evaluation infrastructure. Finally, any arising copyright and licensing issues of datasets created by the PROMISE activities will be managed by ELDA.

3 Outcomes of evaluation activities: CLEF 2010 labs test collections

One of the most important outcomes of benchmarking activities in the context of evaluation campaigns is the test collections they generate. These standardised evaluation resources are crucial for advancing research since they enable meaningful and reproducible comparisons among different approaches, algorithms, theories, and models on common datasets. Evaluation campaigns greatly benefit researchers by mitigating the initial (substantial) effort required for building such test collections. To obtain more information and to gain further insights on the test collections generated by the CLEF 2010 labs, a second questionnaire (see Appendix I: Questionnaires sent to CLEF 2010 Labs organizers) was prepared and sent to the CLEF 2010 lab organizers (together with the one mentioned in Section 2) in May 2011. The results of the two questionnaires regarding the created test collections are presented below and aim to form a point of reference for monitoring the evolution and progress of the CLEF labs over the coming years.

3.1 Collections

The CLEF 2010 Labs employed a total of 13 collections; a description of each collection and some statistics are presented in Appendix IV: CLEF 2010 Labs Test Collections.

All the collections have been purpose-built for the labs and are based on recent crawls. Seven of them were employed for the first time in 2010, while the rest have been used once or at most twice before in previous years of the same labs. These previously used collections have either remained unchanged over these couple of years or, in most cases, they have been gone through substantial updates mainly through the addition of new documents.

About half of the collections are multilingual, ranging from two to nine languages. The monolingual collections include two of the ImageCLEF collections, given that they focus on multimedia retrieval and its language independent nature, and the WePS and PAN-WVC-10 collections that consist of content extracted from web services, such as Twitter and Wikipedia, where English is the dominant language.

The size of the collections and the number of documents they contain vary widely, but the overall trend appears to be towards larger collections with a size of few gigabytes being the norm.

The collections described in this section are evidence of the large size of the datasets already employed in the PROMISE evaluation activities. The continuous update of existing datasets manifests a tendency to increase the volume of data. One of the objectives of PROMISE is to further support this tendency by providing the infrastructure to handle such data. The success of PROMISE in fulfilling this objective will be measured over the coming years by monitoring the changes in indicators such as those reported in the section, e.g., the number of the created data sets and the size of the collections.

3.2 Topics

The nature and the number of topics employed in the tasks of the CLEF 2010 labs depend on the type of the task and are described in Table 12 in Appendix IV: CLEF 2010 Labs Test Collections.

In the classification tasks, the documents to be classified range from 2,000 to around 32,000, while the classes range from 2, to 9, to 93, to a few hundred in the case of CLEF-IP. The number of classes is determined not only based on the requirement of making the tasks realistic, but also on the effort required for generating the ground truth. The same constraints also apply when developing the topics for the rest of the tasks, where retrieval tasks range between 30 topics for the case of highly specialized and domain-specific medical image retrieval and a few thousand topics for the Plagiarism Detection task, while the other tasks have 60-300 topics. It is worth noting that the majority of the tasks employ multilingual topics even if the target collections are monolingual.

Topic creation is an important step in the evaluation campaign cycle and is accompanied by significant challenges in not only creating topics that reflect realistic user information needs, but that these topics are also scientifically feasible and challenging at the same time. The number of topics to be created in the context of an evaluation task is crucial in ensuring the reliability of the experimental outcomes, but is ultimately determined by the effort required in creating the ground truth, as will be discussed next.

3.3 Ground truth

Ground truth generation is one of the major bottlenecks in scaling up the size of test collections given its handicraft nature and required human effort. Table 13 in Appendix IV: CLEF 2010 Labs Test Collections briefly presents the process for the ground truth generation followed in each of the CLEF 2010 tasks and also provides estimates on the applied human effort.

Out of the 13 tasks in the CLEF 2010 Labs, one (LogCLEF) was an exploratory task for which no ground truth was generated, four exploited existing annotations in their collections to automatically generate relevance assessments, whereas the remaining eight tasks employed human assessors. For the latter case, there is a clear trend among these eight tasks to employ crowdsourcing for creating the human relevance assessments, with half (four) of these tasks actually employing such services and in particular those of Amazon's Mechanical Turk. The other half enlisted the help of 3-12 human assessors, mostly volunteers, e.g., students, task organizers, or even task participants, apart from the medical image retrieval task that recruited medical doctors given the specialized nature of the domain of the task.

In the case of automatically generated relevance assessments, ground truth exists for all the documents in the collection. In the case of human relevance assessments, this depends on the size of the collection: for smaller collections, all documents are judged, whereas for larger collections, judging is only applied to pools of top-ranked results; such pools have depths of up to 100 documents. The human effort required to generate these relevance assessments varies greatly based on the nature and difficulty of the task, but can reach up to 300 hours for a single task.

It is clear by the evidence presented in this section that ground truth creation is one of the steps in the evaluation campaign that will benefit tremendously from the automation in the experimental evaluation process currently being investigated by PROMISE. The effects and impact of this automation will become visible in the coming years when adopted by the tasks in the CLEF Labs.

4 Outcomes of the evaluation activities for the “Visual Clinical Decision Support” Use Case

Medicine is one of the most information-intensive fields and potentially affects all of us. Out of all medical exams, imaging has created the largest amount of data available to physicians often with great benefit, but also with a risk of data overload. Finding the right information and making it available to the right persons at the right moment is a challenge. Medical literature currently constitutes an enormous knowledge base that includes visual as well as textual information. Multilingual aspects equally play an important role in this domain as many people are more familiar with formulating information needs in their mother tongue even if they are understanding and speaking English, the language of most of the literature, well. The “Visual Clinical Decision Support” use case aims to analyse the quality that current retrieval technologies deliver on retrieval from the medical literature in several languages and more particularly how visual information analysis can be integrated into the process in the best possible way.

The evaluation activities for this use case take place within the medical retrieval task of the ImageCLEF lab, a task that was organized for the seventh time in 2010. The collection in 2010 contains a total of 77,506 images and captions from the Radiology and Radiographics journals published by RSNA (Radiological Society of North America). This collection constitutes an important body of medical knowledge from the peer-reviewed scientific literature including high quality images with textual annotations. Images are associated with journal articles, and can also be part of larger figures. Figure captions were made available to participants, as well as the sub-caption concerning a particular subfigure (if available). This high-quality set of textual annotations enabled textual searching in addition to content-based retrieval. Furthermore, the PubMed IDs of each figure’s originating article were also made available, allowing participants to access the MeSH (Medical Subject Headings) index terms assigned by the National Library of Medicine for MEDLINE.

Three sub-tasks were conducted by the medical task: *medical modality detection*, *medical image retrieval*, and *medical case retrieval*. The number of registrations to the medical task increased to 51 research groups. However, groups submitting runs have remained stable at 16, with the number of submitted runs increasing to 155. Of these, 61 were image-based retrieval runs, 48 were case-based retrieval runs, while the remaining 46 were modality classification runs.

4.1 Medical Modality Detection Task

The goal of the medical modality detection task is to detect the acquisition modality of the images in the collection. This task is conceived as the first step for the medical image retrieval task, whereby participants use the modality classifier to improve retrieval precision. For this task, 2,390 images were provided as a training set, where each image was classified as belonging to one of 8 classes (CT, GX, MR, NM, PET, PX, US, XR), and 2,620 images were provided as a test set. Each of the images in the test set was to be assigned a modality using visual, textual or mixed techniques. Participants were also requested to provide a classification for all images in the collection. A majority vote classification for all

images in the collection was made available upon request to participants of the task after the evaluation.

A variety of commonly used image processing techniques, classifiers, textual approaches, and their combinations were explored by the participants. Table 2 presents the top-10 results per run type (textual, visual, or mixed). The best results were obtained using mixed methods (94%), while the best run using textual methods (90%) had a slightly better accuracy than the best run using visual methods (87%). However, for groups that submitted runs using different methods, the best results were obtained when they combined visual and textual methods. Further details can be found in [Müller et al., 2010].

Table 2: Top-10 results per run type for the 2010 ImageCLEF Medical Modality Detection task.

Run	Group	Run type	Classification
			Accuracy
XRCE MODCLS COMB testset.txt	XRCE	Mixed	0.94
XRCE MODCLS COMB allset.txt	XRCE	Mixed	0.94
Modality combined.txt	RitsMIP	Mixed	0.93
result text image combined.dat	ITI	Mixed	0.92
result text image comb Max.dat	ITI	Mixed	0.91
result text image comb Prod.dat	ITI	Mixed	0.91
gigabioinformatics-both.txt	GIGABIOINFORMATICS	Mixed	0.90
result text image comb CV.dat	ITI	Mixed	0.89
result text image comb Sum.dat	ITI	Mixed	0.87
Modalityall Mix.txt	RitsMIP	Mixed	0.78
XRCE MODCLS TXT allset.txt	XRCE	Textual	0.90
result text titile caption mod Mesh.dat	ITI	Textual	0.89
entire text based modality class.dat	ITI	Textual	0.86
gigabioinformatics-text.txt	GIGABIOINFORMATICS	Textual	0.85
Modality text.txt	RitsMIP	Textual	0.85
Modalityall Text.txt	RitsMIP	Textual	0.85
ipl aueb rhcpp full CT.txt	AUEB	Textual	0.74
ipl aueb rhcpp full CTM.txt	AUEB	Textual	0.71
ipl aueb rhcpp full CTMA.txt	AUEB	Textual	0.53
ipl aueb svm full CT.txt	AUEB	Textual	0.53
XRCE MODCLS IMG allset.txt	XRCE	Visual	0.87
UESTC modality boosting	UESTC	Visual	0.82
UESTC modality svm	UESTC	Visual	0.80
result image comb sum.dat	ITI	Visual	0.80
result image comb CV.dat	ITI	Visual	0.80
entire result image comb CV.dat	ITI	Visual	0.80
entire result image comb CV.dat	ITI	Visual	0.80
result image combined.dat	ITI	Visual	0.79
entire result image combined.dat	ITI	Visual	0.79
result image comb Max.dat	ITI	Visual	0.76

4.2 Medical Image Retrieval Task

The goal of the image-based medical retrieval task is to retrieve a ranked set of images that best meet an information need specified as a textual statement and a set of sample images. Realistic information needs were identified by conducting a user study at Oregon Health & Science University (OHSU) that focused on understanding the needs of medical practitioners, both met and unmet, regarding medical image retrieval. The participants in this user study provided textual descriptions of their information needs in English. In 2010, 16 of them were selected as topics for the medical image retrieval task. These topics were translated to French and to German, and 2 to 4 sample images were added to each. This set of topics was also approved by a physician. Relevance judgements were performed with the same on-line system as in 2008 and 2009. Judges were provided with a protocol for the process with specific details on what should be regarded as relevant versus non-relevant. A ternary relevance judgement scheme was used, wherein each image in the pool was judged to be “relevant”, “partly relevant”, or “non-relevant”. Judges were recruited by sending out an e-mail to current and former students at OHSU’s Department of Medical Informatics and Clinical Epidemiology. Judges, primarily clinicians, were paid a small stipend for their services.

The best results for the ad-hoc retrieval topics were obtained using mixed methods. Textual methods also performed well, but visual methods by themselves, were not very effective for this collection. Table 3 presents the top-10 results per run type (textual, visual, or mixed): only 8 of the 61 submitted runs used purely visual techniques. Given that this collection contains extremely well annotated textual captions and images that are primarily from radiology, it does not lend itself to purely visual techniques. However, as seen from the results of the mixed runs, the use of the visual information contained in the image can improve the search performance over that of a purely textual system. Participants explored a variety of textual retrieval techniques and many found the use of the manually assigned MeSH terms to be most useful. Modality filtration, using either text-based or image-based modality detection techniques was found to be useful by some participants while others found only minimal benefit using the modality. The run with the highest MAP utilized a multimodal approach to retrieval. However, many groups that performed a pure fusion of the text-based and image-based runs found a significant deterioration in performance as the visual runs had very poor performance. This year’s results again emphasize the previously noted observations that although the use of visual information can improve the search results over purely textual methods, the process of effectively combining the information from the captions and image itself can be quite complex and are often not robust. Simple approaches of fusing visual and textual runs rarely lead to optimized performance. Further details can be found in [Müller et al., 2010].

Table 3: Top-10 results per run type for the 2010 ImageCLEF Image-based Medical Retrieval task.

Run	Run type	Group	MAP	bPref	P10
XRCE AX rerank comb.trec	Mixed	XRCE	0.3572	0.3841	0.4375
XRCE CHI2 LOGIT IMG MOD late.trec	Mixed	XRCE	0.3167	0.3610	0.3812
XRCE AF LGD IMG late.trec	Mixed	XRCE	0.3119	0.3201	0.4375
WIKI AX IMG MOD late.trec	Mixed	XRCE	0.2818	0.3279	0.3875
OHSU all mh major all mod reorder.txt	Mixed	OHSU	0.2560	0.2533	0.3813
OHSU high recall.txt	Mixed	OHSU	0.2386	0.2533	0.3625
queries terms 0.1 Modalities.trec	Mixed	ITI	0.1067	0.1376	0.2812
XRCE AX rerank.trec	Mixed	XRCE	0.0732	0.1025	0.1063
Exp Queries Cit CBIR CV MERGE MAXt	Mixed	ITI	0.0641	0.0962	0.1438
runMixt.txt	Mixed	UAIC2010	0.0623	0.0666	0.1313
WIKI AX MOD late.trec	Textual	XRCE	0.3380	0.3828	0.5062
ipl aueb AdHoc default TC.txt	Textual	AUEB	0.3235	0.3109	0.4687
ipl aueb adhoq default TCg.txt	Textual	AUEB	0.3225	0.3087	0.4562
ipl aueb adhoq default TCM.txt	Textual	AUEB	0.3209	0.3063	0.4687
ipl aueb AdHoc pivoting TC.txt	Textual	AUEB	0.3155	0.2998	0.4500
ipl aueb adhoq Pivoting TCg.txt	Textual	AUEB	0.3145	0.2993	0.4500
ipl aueb adhoq Pivoting TCM.txt	Textual	AUEB	0.3102	0.3005	0.4375
OHSU pm all all mod.txt	Textual	OHSU	0.3029	0.3440	0.4313
OHSU pm major all mod.txt	Textual	OHSU	0.3004	0.3404	0.4375
OHSU all mh major jaykc mod.txt	Textual	OHSU	0.2983	0.3428	0.4188
fusion cv merge mean.dat	Visual	ITI	0.0091	0.0179	0.0125
XRCE IMG max.trec	Visual	XRCE	0.0029	0.0069	0.0063
fusion cv merge max.dat	Visual	ITI	0.0026	0.0075	0.0063
GE GIFT8.treceval	Visual	medGIFT	0.0023	0.0060	0.0125
NMFAsymmetricDCT5000 k2 7	Visual	Bioingenium	0.0018	0.0110	0.0063
fusion cat merge max.dat	Visual	ITI	0.0018	0.0057	0.0063
NMFAsymmetricDCT2000 k2 5	Visual	Bioingenium	0.0015	0.0079	0.0063
NMFAsymmetricDCT5000 k2 5	Visual	Bioingenium	0.0014	0.0076	0.0063

4.3 Medical Case Retrieval Task

The goal of the case-based medical retrieval task is to return a ranked set of articles (rather than images) that best meet the information need provided as a description of a “case”. The aim is to move medical retrieval potentially closer to clinical routine by simulating the use case of a clinician who is in the process of diagnosing a difficult case. Providing clinicians with articles from the literature that discuss cases similar to the case they are working on

can be a valuable aid to choosing a good diagnosis or treatment.

Fourteen topics were created based on cases from the teaching file Casimage that contains cases (including images) from radiological practice that clinicians document mainly for use in teaching. The diagnosis and all information on the chosen treatment were removed from the cases so as to simulate the situation of the clinician who has to diagnose the patient. In order to make the judging more consistent, the relevance assessors were provided with the original diagnosis for each case. The relevance judgements were performed in the same manner as for the medical image retrieval task (see Section 4.2) by adapting the system for the case-based topics.

Table 4: Top-10 results per run type for the 2010 ImageCLEF Case-based Medical Retrieval task.

Run	Run type	Group	MAP	bPref	P10
PhybaselineRelfbWMR 10 0.2sub	Textual Feedback	UIUCIBM	0.3059	0.3348	0.4571
PhybaselineRelfbWMD 25 0.2sub	Textual Feedback	UIUCIBM	0.2837	0.3127	0.4571
PhybaselineRelfbWMR 10 0.2 top20sub	Textual Feedback	UIUCIBM	0.2713	0.2897	0.4286
case queries pico backoff 0.1	Textual Feedback	ITI	0.1386	0.1666	0.2000
PhybaselinefbWMR 10 0.2sub	Textual Manual	UIUCIBM	0.3551	0.3714	0.4714
PhybaselinefbWsub	Textual Manual	UIUCIBM	0.3441	0.3480	0.4714
PhybaselinefbWMD 25 0.2sub	Textual Manual	UIUCIBM	0.3441	0.3480	0.4714
Case expanded queries terms 0.1	Textual Manual	ITI	0.0601	0.0825	0.0857
baselinefbWMR 10 0.2sub	Textual Automatic	UIUCIBM	0.2902	0.3049	0.4429
baselinefbWsub	Textual Automatic	UIUCIBM	0.2808	0.2816	0.4429
hes-so-vs case-based fulltext.txt	Textual Automatic	HES-SO	0.2796	0.2699	0.4214
baselinefbsub	Textual Automatic	UIUCIBM	0.2754	0.2856	0.4286
baselinefbWMD 25 0.2sub	Textual Automatic	UIUCIBM	0.2626	0.2731	0.4000
C TA T.lst	Textual	SINAI	0.2555	0.2518	0.3714
IRIT SemAnnotator-2.0 BM25 N28	Textual Automatic	IRIT	0.2265	0.2351	0.3429
C TA TM.lst	Textual	SINAI	0.2201	0.2307	0.3643
IRIT SemAnnotator-2.0 BM25 N28 1	Textual Automatic	IRIT	0.2193	0.2139	0.3286
IRIT SemAnnotator-1.5.2 BM25 N34	Textual Automatic	IRIT	0.2182	0.2267	0.3571
GE GIFT8 case	Visual Automatic	medGIFT	0.0358	0.0612	0.0929
case queries cbir with case backoff	Mixed Automatic	ITI	0.0353	0.0509	0.0429
case queries cbir without case backoff	Mixed Automatic	ITI	0.0308	0.0506	0.0214
GE Fusion case captions Vis0.2	Mixed Automatic	medGIFT	0.0143	0.0657	0.0357
GE Fusion case fulltext Vis0.2	Mixed Automatic	medGIFT	0.0115	0.0786	0.0357

Almost all groups focused on using textual retrieval techniques, as combining visual retrieval on a case basis is a difficult approach. Table 4 presents the top-10 results per run type

(textual (automatic or feedback), visual, or mixed): only 1 run used purely visual techniques, while only two participants submitted a total of 4 mixed runs. In addition, there were actually a substantial number of feedback textual runs. Best results were obtained with a textual retrieval approach when using relevance feedback. The performance of the single visual run submitted shows that the results are much lower than the text-based techniques. Still, compared with the image-based retrieval only a single image-based run had a higher MAP, meaning that also case-based retrieval is possible with purely visual retrieval techniques and can be used as a complement to the text approaches. The first three textual automatic runs are basically very close and then the performance slowly drops. In general results are slightly lower than for the image-based topics. Relevance feedback can improve results, although the improvement is fairly low compared with the automatic run. All but one of the feedback runs had very good results, showing that the techniques work in a stable manner. The performance of the mixed runs is fairly low, highlighting the difficulty in combining the textual and visual results properly. Much more research on the visual and combined retrieval seems necessary as the current techniques in this field do not seem to work in a satisfying way. Further details can be found in [Müller et al., 2010].

4.4 Summary of the outcomes of the “Visual Clinical Decision Support” Use Case

The main outcomes of the first year evaluation activities for the “Visual Clinical Decision Support” use case realised within the medical retrieval task at ImageCLEF are:

1. The task remains popular, even in its seventh edition, attracting a high number of registrations and participations. There was an increase in the number of submitted runs (155), the highest number of experiments submitted in any of the CLEF Labs.
2. Combination of evidence from different modalities is the most effective approach for the modality detection and medical image retrieval tasks, whereas further research is needed for the medical case retrieval task. In particular:
 - a. For the modality detection task, although textual and visual methods alone were relatively successful, combining these techniques proved most effective. Furthermore, groups that submitted runs using different methods obtained their best results when they combined visual and textual methods.
 - b. For the medical image retrieval task, the best results were obtained using mixed methods, with textual methods also performing well, but with visual methods not being very effective by themselves. However, many groups that performed a pure fusion of the text-based and image-based runs found a significant deterioration in performance as the visual runs had very poor performance. In conclusion, the process of effectively combining the information from the captions and image itself can be quite complex and often not robust. Simple approaches of fusing visual and textual runs rarely lead to optimized performance.
 - c. For the medical case retrieval task, textual methods were clearly superior. The performance of the single visual run submitted shows that the results are much lower than the text-based techniques. The performance of the mixed runs is fairly low, highlighting the difficulty in combining the textual and visual

results properly. Much more research on the visual and combined retrieval seems necessary as the current techniques in this field do not seem to work in a satisfying way.

3. Interactive retrieval is still being used only by a very small number of participants, although it does have the potential to improve retrieval effectiveness.
 - a. For the medical case retrieval, there were actually a substantial number of feedback textual runs. Best results were obtained with a textual retrieval approach when using relevance feedback.
 - b. To encourage research in this area, a medical user-oriented (interactive) image retrieval task is organized for CLEF 2011.
4. A kappa analysis between several relevance judgements for the same topics shows that, although there are differences between judges, there was agreement on topics that have more than 10 relevant images. The relative rankings of the groups were vastly unchanged with using the assessment of different judges aside from topics with low number of relevant images. As a result, topics with very few relevant images could be removed or a more thorough testing could already remove them during the topic creation process.

5 Outcomes of the evaluation activities for the “Search for Innovation” Use Case

Searching for innovation in technological areas involves searching in patent collections for assessing the state-of-the-art on a technical subject at a given point in time. The evaluation activities for the “Search for Innovation” Use Case centre on the CLEF-IP lab, a benchmarking activity on intellectual property, that aims at evaluating patent retrieval. In 2010, CLEF-IP organized two tasks corresponding to two important steps in the process: *Prior Art Candidates Search* (PAC) and *Patent Classification* (CLS). Table 5 lists the 12 groups that participated in the two tasks, submitting 25 runs to the PAC task and 27 runs to the CLS task.

Table 5: List of participants to CLEF-IP 2010 and runs submitted to the CLS and PAC tasks.

Group ID	Institution		CLS	PAC	PAC topic set
bitem	BiTeM, Service of Medical Informatics, Geneva University Hospitals	CH	7	2	large
dcu	Dublin City Univ. - School of Computing	IE		3	large
hild	Hildesheim Univ. - Information Science	DE		4	small
humb	Humboldt Univ. - Dept. of German Language and Linguistics	DE	1	1	large
insa	LCI - Institut National des Sciences Appliquées de Lyon	FR	5		
jve	Industrial Property Documentation Department, JSI Jouve	FR	3		
run	Information Foraging Lab, Radboud University Nijmegen	NL	2	2	small
spq	Spinque	NL	1	1	large
ssft	Simple Shift	CH	8		
uaic	Al. I. Cuza University of Iasi, Natural Language Processing	RO		1	large
ui	Information Retrieval Group, Universitas Indonesia	ID		3	large
uned	UNED - E.T.S.I. Informatica, Dpto. Lenguajes y Sistemas Informaticos	ES		8	large
Total Runs			27	25	

Sections 5.1 and 5.2 present the outcomes of these two tasks, respectively, while Section 0 discusses the lessons learned in the PatOlympics and the TREC Chemical IR Evaluation campaign, two additional evaluation activities that have benefited the “Search for Innovation” Use Case. Section 5.4 concludes by summarizing the outcomes for the “Search for Innovation” Use Case in this first year of evaluation activities.

5.1 Prior Art Candidates Search Task

The objective of the Prior Art Candidates search task is to retrieve documents from a collection of patents that could constitute prior art for a given topic patent. In this context, a patent's "prior art" refers to patents with technical details similar to some or all technical details described in the given patent.

In 2010, the topic set consists of 2,000 topics, while participants with less computing power were allowed to submit runs for a smaller subset of 500 topics. Each run consisted of a single text file with at most 1,000 answers per topic. The relevance assessments were automatically extracted from the patent expert search reports. These reports list relevant patent documents that the patent experts found during their prior art searches.

Table 5 lists the 9 teams that submitted runs; in total, 25 runs were submitted, 6 runs with retrieval experiments for the small set of topics, and 19 for the large set of topics. We created two evaluation bundles corresponding to the small and large experiment sizes: (i) a small evaluation bundle, and (ii) a large evaluation bundle.

The large evaluation bundle contains all runs that returned results for the large topic set. The small evaluation bundle contains all runs with results for the small topic set, together with the runs submitted to the large topic set after restricting their content to the small topic set. In addition to the two size bundles, evaluations were also done on each of three topic sets where the topic document language was English, German, or French, respectively. These topic sets were extracted from the large topic set, resulting in 1,348 English language topics, 518 German topics, and 134 French topics. Only the runs using the large topic set were included in these topic language evaluation bundles.

For each submitted run, the following measures were computed:

- ⤴ Precision, Precision@5, Precision@10, Precision@50, Precision@100
- ⤴ Recall, Recall@5, Recall@10, Recall@50, Recall@100
- ⤴ Map
- ⤴ nDcg
- ⤴ PRES¹

The retrieval methods involved by the participants to this task range from out-of-the box, configurable retrieval systems, like Lucene and Indri, to retrieval systems enriched with dictionaries and multi-lingual support. Since topics were complete patent documents (XML files), participants were given freedom in creating retrieval queries. The range of methods for query generation varies not only in the algorithms used to extract the query terms (like tf-idf based and language model based) but also in the part of the document that was used to create the query. Generally, retrieval systems that included patent specific meta-data, stored in the collection documents, along with text processing algorithms performed better than the ones that did not include it.

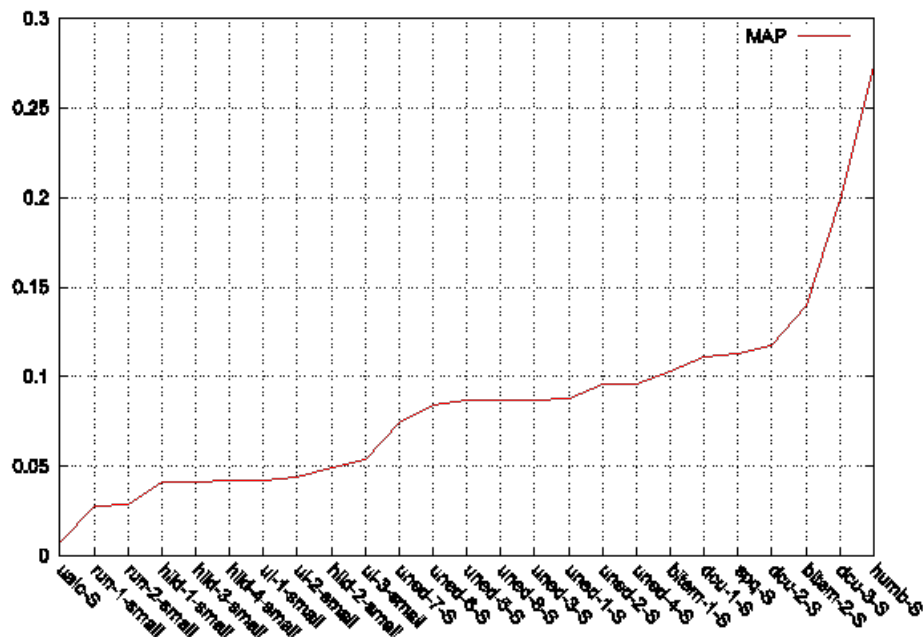
The multi-lingual aspect of the PAC topics was exploited in a rather modest fashion, most of the participants choosing to translate queries using Google Translate. The generally retrieval

¹ Magdy W. and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In SIGIR 2010

results did not improve significantly. This may actually reflect the fact that English language documents are overrepresented in the CLEF-IP collection. We must also remark that the language distribution over the PAC topics was not ideal, with English being over-represented. This is remedied for in the 2011 CLEF-IP Lab.

Figure 2 shows the plotted MAP values for the small evaluation bundle. For all other computations, see [Piroi 2010a].

Figure 1: MAP measures for the 2010 CLEF-IP PAC small evaluation bundle.



5.2 Patent Classification Task

The objective of the second task in CLEF-IP 2010 was to classify a given patent document according to the International Patent Classification system (IPC²). The classification was to be given at the subclass level. The set of classification topics contained 2,000 patent documents, a different set than the one used in the PAC task. The relevance assessments were automatically extracted from the classification information stored in the document that originated the classification topics.

The measure computations were made for four sets of topics: the complete set of classification topics, and the three language-based topic subsets. Out of the 2,000 classification topics, there were 1,470 topics with English as the document language, 408 with German as the document language, and 122 topics with French as the document language.

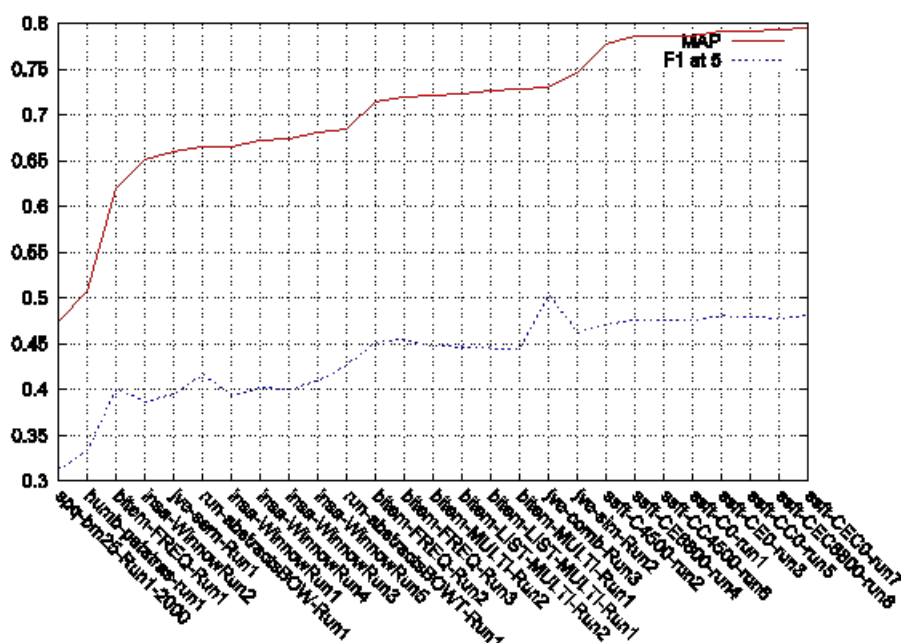
The approaches used to classify the set of topics vary from well-known classification

² The IPC system is maintained by the World Intellectual Property Organization, and is a classification system hierarchically organized in sections, classes, subclasses, and groups.

algorithms (kNN, Winnow-based) to combining retrieval and classification algorithms. The approaches where the topic's IPC subclass code is extracted from documents considered relevant to the classification topic do not perform as well as the well-known classification algorithms. The language of the topic document did not seem to impact on the classification results. As in the PAC case, we remark that the language distribution over the CLS topics was not ideal, with English being over-represented. This is remedied for in the 2011 Lab.

Figure 3 shows the plotted MAP and F1@5 measure values for the complete set of topics. For all other computations see [Piroi 2010b].

Figure 2: MAP and F1 measures for the 2010 CLEF-IP CLS task.



5.3 Other activities for the “Search for Innovation” sse case

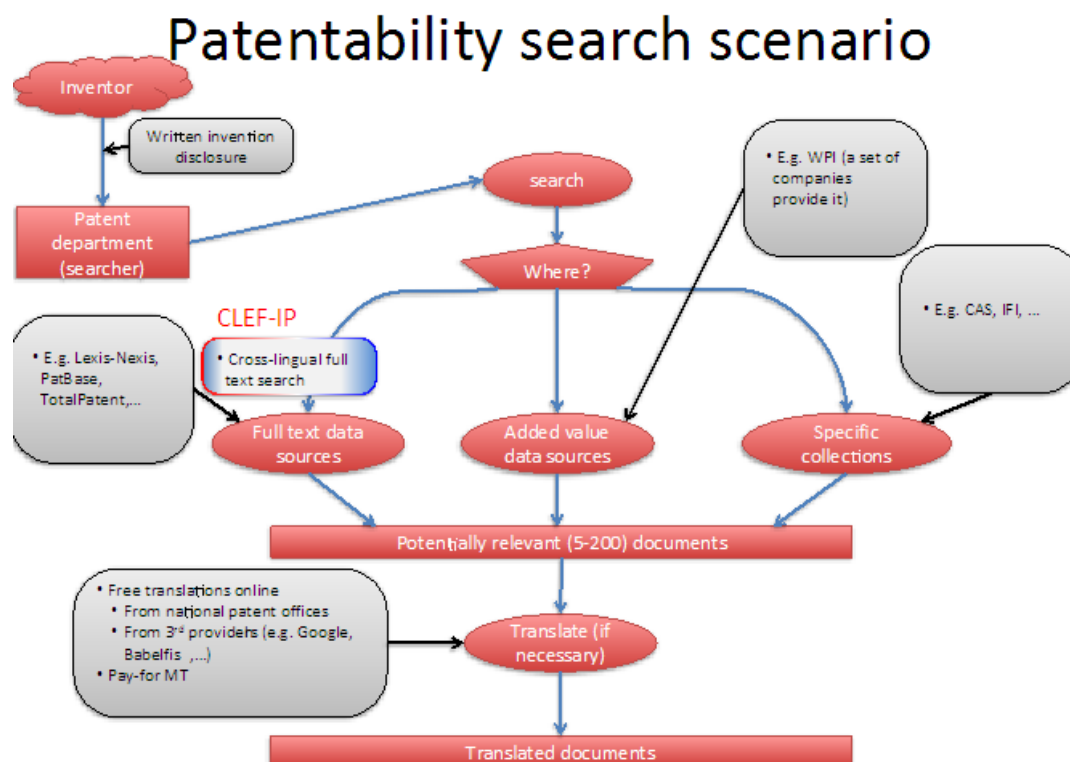
In addition to the CLEF-IP Lab, the “Search for Innovation” Use Case has benefited from lessons learned in the PatOlympics and the TREC Chemical IR Evaluation campaign in which the IRF/TUWien team has also been involved.

PatOlympics is a cross-domain interactive workshop that helps us, via a direct interaction with the users, better understand their needs and consequently model the CLEF-IP labs more appropriately. It is particularly useful for the participants to see exactly how the users search, in order to devise better systems for CLEF-IP. In 2011, PatOlympics had 5 participants, 4 of which participate also in CLEF-IP or in TREC-CHEM or both.

Discussions with IP specialists on the issue of cross-lingual search have resulted in the definition of a workflow described in Figure 3. While it is true that a professional searcher will use several tools to search for patents, full text search has become more and more required and this is what the lab specializes in. At the same time, the patent specialists point out the need to understand that different domains lay different weights on the different

kinds of search tools. Cross-lingual search is for instance less important for chemical patents, because of the universal language of chemistry. Chemical patents remain however the single largest category of patents and are considered of paramount importance.

Figure 3: Patentability search workflow in a multi-lingual field



5.4 Summary of the outcomes of the “Search for Innovation” use case

A number of important lessons were learned from the experience of the evaluation activities for the “Search for Innovation” Use Case in this first year of evaluation activities. Some of the problems identified have already been addressed in the 2011 CLEF-IP lab, as well as in a stronger collaboration with other evaluation efforts. The following list summarizes them:

1. The CLEF-IP Lab does not evaluate the full “search for innovation” process, as understood by the professional patent searchers:
 - a. This is normal and agreed: CLEF-IP focuses on the cross-lingual full text retrieval evaluation. A better communication to professional users is needed to convey the objectives of the Lab.
 - b. Any “search for innovation” performed by professional patent searches is done in rounds, with continuous refining of the query. This is not currently reflected in the CLEF-IP Lab.

2. The “Search for innovation” is very domain specific. Chemistry, for example, has a retrieval process on its own and cross-lingual search is useless in such a domain where there exists a universal language:
 - a. We collaborate with NIST on the organization of TREC-CHEM and try to bring that evaluation campaign into the Network of Excellence.
3. Classification at IPC class level performs well across most participating groups, indicating that the problem may be trivial:
 - a. Lab participants that have a commercial background performed clearly better than participants from academic institutions.
 - b. In 2011, we introduced a task for IPC classification at sub-group level.
4. Users find the patent search hard to understand and lack motivation to participate
 - a. We need to involve the users more directly and show them how each system works. In this sense, PROMISE supports the PatOlympics evaluation campaign.

6 Outcomes of the evaluation activities for the “Unlocking Culture” Use Case

6.1 CHiC2011

The “Unlocking Culture” domain deals with effective information access to cultural heritage material held in large-scale digital libraries containing data from libraries, archives, museums, and audio-visual archives. However, access to these objects still poses several challenges related to the heterogeneous media types (texts, audio or video files) and user groups (novice and expert users), often with specialized information needs. Objects are provided by metadata, usually in their national language and with specified technical vocabularies suited for their particular domains. Research within this domain focuses on the satisfaction of user information needs by retrieving relevant “cultural assets” irrespective of the media type, location or language in which information objects are expressed. Even though digital libraries are constantly growing and much research is carried out in the field, much less is done to establish standard evaluation criteria and methods.

The **CHiC2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage** workshop at CLEF 2011 aims to investigate these issues by surveying evaluation efforts in the cultural heritage (CH) field as well as defining user scenarios and identifying possible relevant metrics (<http://www.promise-noe.eu/chic-2011/home>). The workshop provides an overview of previous or current evaluation activities and seeks to introduce an exchange about future efforts that needs to be addressed in the CH field. The objective of this workshop is to review existing use cases in the CH domain and their translation into potential retrieval and evaluation scenarios that can be used as benchmarks for evaluating CH information access systems.

The overall goals are:

1. To establish what makes searching in the CH domain distinct from other domains.
2. To gather existing use cases for multilingual information access in the CH domain.
3. To review existing evaluation resources studies within the CH domain.
4. To propose appropriate methodologies for evaluating multilingual information access to CH resources.
5. To define multiple concrete evaluation tasks modeled on IR evaluation initiatives such as CLEF, TREC or INEX.

Invited talks will address use cases, evaluation approaches, and best practices from an institutional point of view as well as the experiences from large-scale evaluation campaigns. Participants are asked to contribute statements concerning complementary efforts, projects, initiatives and available test data. Based on the speakers input and group discussions the second part of the workshop aims to identify possible synergies between evaluation frameworks within CH projects and campaigns as well as the development of use cases and usage scenarios that can be applied to CH information systems.

The keynote speakers are:

- ✧ Jaap Kamps <http://staff.science.uva.nl/%7Ekamps/> is an Assistant Professor of Information Retrieval at the Faculty of Humanities, University of Amsterdam. He was involved in several externally funded research projects such as MuSeUM (Multiple-collection Searching Using Metadata) and MultiMATCH (Multilingual/Multimedia Access To Cultural Heritage).
- ✧ Johan Oomen <http://www.linkedin.com/in/johanoomen> is head of the Netherlands Institute for Sound and Vision R&D Department and researcher at the Web and Media group of the Vrije Universiteit Amsterdam. His research focuses on providing access to digital heritage on the Web such as Europeana V1.0 or PrestoPRIME.
- ✧ Christos Papatheodorou <http://www.ionio.gr/%7Epapatheodor/> is an Associate Professor at the Department of Archives and Library Sciences, Ionian University, Corfu, Greece and a fellow researcher in the Digital Curation Unit, Institute for the Management of Information Systems, "Athena" Research Centre, Athens, Greece. He was the general co-chair of the 13th ECDL and organized the tutorial "Exploring Perspectives on the Evaluation of Digital Libraries".

6.2 Reports: State-of-the-Art Evaluation of Digital Libraries in the Cultural Heritage domain

Digital libraries and other information systems that access cultural heritage (CH) content are becoming increasingly complex as they manage various content input from different CH institutions, e.g., libraries, museums and archives, and provide access to that material in a unified and coherent way coping with different media types (text, sound, image and video), data models and multilingual aspects of the descriptions and the content of the stored items/objects. For the "Unlocking Culture" use case and as preparation for the CHiC-workshop we prepare two overview reports describing the state-of-the-art of evaluation of digital libraries in the cultural heritage-domain. We follow a two-pronged approach to present the state-of-the-art: one report describes the important use case components in existing CH information systems and the other one analyzes current evaluation approaches for CH information systems. Next, we present an outline of the two reports.

6.2.1 Report on Use Case Components in Cultural Heritage Information Systems

Since PROMISE applies a use case-based approach for the planned evaluation activities, the first report deals with use cases/ use case components that refer to the CH domain. The method of data gathering was the review and analysis of about 25 CH projects and 10 evaluation campaigns with the specific perspective on which use cases/components/scenarios have been identified or worked with by CH information systems.

As CH information systems are heterogeneous, various scenarios of usage are possible. Our work includes a generalization or abstraction from individual scenarios to use case components for the "Unlocking culture" domain, therefore establishing building blocks for interaction scenarios. We distinguish between three possible patterns of user interactions for the CH domain: *Search*, *Explore and Discover*, and *Engage*. Search deals with all

interactions that require a user to actively input a query to the system. Explore and Discover are characterized by other information gathering interactions of users for content or by getting information the CH information system itself. Engage comprises all interactions by users that refer to contributing content to CH information systems.

6.2.2 Report on Evaluation in Cultural Heritage

The second report gives an overview on evaluation methods for CH information systems that have been applied so far by cultural heritage projects and or in the context of large-scale evaluation campaigns. The method of data gathering was the review and analysis of about 25 CH projects and 10 evaluation campaigns with the specific perspective on which evaluation methods have been applied or established so far in the evaluation of information systems in the CH domain.

The report deals with evaluation approaches for CH information systems from two perspectives: the *system-centric* and the *user-centric* perspective. Cultural heritage systems have served as test cases and provided data sets to large-scale international information retrieval evaluation initiatives, for example in the TEL (The European Library) track at CLEF or the INEX book track. Thus the evaluation campaigns also developed test collections and metrics that are relevant for evaluation in CH. We describe system-centric evaluation methods that refer to laboratory-based evaluation of information systems in terms of information retrieval tests as the standard evaluation approach focusing on effectiveness (using standard IR metrics, e.g. recall and precision). We also present the user-centric evaluation methods (e.g. usability tests, user satisfaction surveys, attitude interviews) in the report, focusing on user behavior, user satisfaction and interface usability. They can be applied to complement the system retrieval performance measurements.

7 Further outcomes of evaluation activities: information and knowledge resources

Besides fostering, supporting, and coordinating experimental evaluation activities, PROMISE also aims to curate, preserve, and enrich the information and knowledge resources resulting from such activities, such as experimental data, methodologies, and publications, as well as the possible relationships among them; the ultimate goal is to provide access to such resources so that they can be put to use by the targeted communities towards the research and development of multimedia and multilingual information access systems. In this first year of PROMISE, we have evaluated the quality and impact of two such resources: (i) the CLEF-derived publications and (ii) the Cross Language Evaluation Forum (CLEF) websites.

The publications derived from specific research activities and the citations (academic references) they receive are commonly used to assess the scientific, and in particular the scholarly, quality and impact of these activities. PROMISE aims to measure the scholarly impact of the evaluation activities of the CLEF campaigns in order to monitor the progress with respect to its objectives. To this end, HES-SO has initiated a study on assessing the scholarly impact of CLEF by conducting a preliminary assessment of the scholarly impact of ImageCLEF, the cross-language image retrieval evaluation initiative that has been running as part of CLEF since 2003. Section 7.1 reports on the main outcomes of this preliminary analysis; further details can be found in [Tsikrika et al., 2011].

The CLEF websites are used to not only support and promote the visibility of the evaluation activities of the CLEF campaigns, but to also archive their outcomes, and thus act a vehicle for the dissemination of information and knowledge on multilingual and multimedia information systems. Currently, the following websites are associated with CLEF: (i) the CLEF campaign website³ which maintains information on the evaluation activities that have been conducted over the first ten years of CLEF from 2000 to 2009, and (ii) the CLEF 2010⁴ and CLEF 2011⁵ websites that provide all the necessary information about CLEF 2010 and 2011, respectively, including information about the CLEF conference and the CLEF labs, as well as electronic versions of the labs notebook papers. One of our objectives is the future integration of these websites with the PROMISE open evaluation infrastructure that aims to support a growing knowledge-base, where experimental collections, experimental results and evidence, evaluation measures and analyses will accumulate and be available for further study. To this end, a qualitative analysis of the CLEF campaign website has been carried out by UNIPD so as to gain some indications on the way users perceive the website. Section 7.2 reports on the outcomes of this analysis; these results are currently being used for redesigning the CLEF campaign website so as to make it more responsive to user requirements and to enable its smooth integration with the PROMISE evaluation

³ <http://www.clef-campaign.org/>

⁴ <http://www.clef2010.org/>

⁵ <http://www.clef2011.org/>

infrastructure. Evidence on the high visibility of the CLEF websites, and in particular of the CLEF 2010 website, is provided in Appendix VI: CLEF 2010 website statistics.

7.1 CLEF campaign: scholarly impact analysis

Recent investigations have reported on the scholarly impact of the TRECVID [Thornley et al., 2011] and on the economic impact of the TREC [Rowe et al., 2010] evaluation campaigns. Building on this work, HES-SO initiated a study on assessing the scientific impact of CLEF. This section presents the main findings of a preliminary study on assessing the scholarly impact of ImageCLEF by performing a citation analysis on a dataset of ImageCLEF-derived publications and focusses particularly on the results pertaining to the medical image retrieval task corresponding to the “Visual Clinical Support Decision” Use Case. Section 7.1.1 describes the bibliometric analysis method applied for assessing the scholarly impact of ImageCLEF, while Section 7.1.2 presents the results of this analysis. Further details can be found in [Tsirikla et al., 2011].

7.1.1 ImageCLEF scholarly impact: bibliometric analysis method

ImageCLEF, the cross-language image retrieval evaluation campaign, was introduced in CLEF 2003 and has organized a number of tasks within two main domains: (i) medical image retrieval, and (ii) general (non-medical) image retrieval from historical archives, news photographic collections, and Wikipedia pages. Table 6 summarizes the ImageCLEF tasks that ran between 2003 and 2010 and shows the number of participants for each task along with the distinct number of participants in each year. The number of participants and tasks in ImageCLEF has continued to grow steadily throughout the years, from four participants and one task in 2003, reaching its peak in 2009 with 65 participants and seven tasks.

Table 6: Participation in the ImageCLEF tasks and number of participants by year.

Task	2003	2004	2005	2006	2007	2008	2009	2010
General images								
Photographic retrieval	4	12	11	12	20	24	19	-
Interactive image retrieval	1	2	2	3	-	6	6	-
Object and concept recognition				4	7	11	19	17
Wikipedia image retrieval						12	8	13
Robot vision							7	7
Medical images								
Medical image retrieval		12	13	12	13	15	17	16
Medical image annotation			12	12	10	6	7	-
Total (distinct)	4	17	24	30	35	45	65	49

Bibliometric studies provide a quantitative and qualitative indication of the scholarly impact of research by examining the number of scholarly publications derived from it and the number of citations these publications receive. The most comprehensive sources for publication and in particular for citation data are: (i) the Thomson Reuters Web of Science (generally known as ISI Web of Science or ISI), established by Eugene Garfield in the 1960s, (ii) Scopus, introduced by Elsevier in 2004, and (iii) the freely available Google Scholar,

developed by Google in 2004. In addition to publication and citation data, ISI and Scopus also provide citation analysis tools to calculate various metrics of scholarly impact, such as the h-index, a robust metric of scientific research output that has a value h for a dataset of N publications, if h of them have at least h citations each, and the remaining $(N-h)$ publications have no more than h citations each. Google Scholar on the other hand is simply a data source and does not have such capabilities; citation analysis using its data can though be performed by the Publish or Perish (PoP) system, a software wrapper for Google Scholar. Given that ISI has a limited coverage of publications in conference proceedings in the field of computer science, this study employs Scopus and Google Scholar (in particular its PoP wrapper) for assessing the scholarly impact of ImageCLEF. This allows us to also explore a further goal: to compare and contrast these two data sources in the context of such an analysis. Scopus and Google Scholar were also employed in the examination of the TRECVID scholarly impact [Thornley et al., 2011], where emphasis was mostly given on the Google Scholar data.

To assess the scholarly impact of ImageCLEF, bibliometric analysis can be applied to the dataset of publications that contains:

- ✧ ImageCLEF-related publications in the CLEF working notes: Although publications in the CLEF working notes do attract citations, given that Scopus does not index them, they are excluded from our analysis, so as to allow a “fair” comparison between the two citation data sources.
- ✧ ImageCLEF-related publications in the CLEF proceedings: These publications are indexed by both Scopus and Google Scholar and therefore are included in our analysis.
- ✧ Papers describing ImageCLEF resources: Given that these publications are written by ImageCLEF organizers, they were located by searching by author name. The results were manually refined by an expert in the field and added to the dataset of publications to be analysed.
- ✧ ImageCLEF-derived publications: Locating all publications that use Image-CLEF data is a hard task. One may assume that such papers would cite the overview article of the corresponding year of ImageCLEF, but often only the URL of the benchmark is mentioned, or that such papers are written by researchers having access to the data. Both such searches in Scopus and PoP require extensive manual data cleaning and the inclusion of such publications in the analysis is left as part of the next stage of our investigation.

Therefore, this preliminary study to assess the scholarly impact of ImageCLEF focusses on the analysis of the dataset of publications published between 2004 and 2010 and consisting of (i) ImageCLEF-related participants' and overview papers in the CLEF proceedings, and (ii) overview papers regarding ImageCLEF resources published elsewhere.

7.1.2 ImageCLEF scholarly impact: results

The results of our study, presented in Table 7, show that there were a total of 195 ImageCLEF-related papers in the CLEF proceedings published between 2004 and 2010. Over the years, there is a steady increase in such ImageCLEF publications, in line with the continuous increase in participation and in the number of offered tasks (see Table 6). The

coverage of publications regarding ImageCLEF resources varies greatly between Scopus and Google Scholar, with the former indexing a subset that contains only 57% of the publications indexed by the latter. These publications peak in 2010, which coincides with the year that ImageCLEF organised a benchmarking activity as a contest in the context of the International Conference for Pattern Recognition (ICPR). This event was accompanied by several overview papers describing and analysing the Image-CLEF resources used in the contest, published in the ICPR 2010 and ICPR 2010 Contests proceedings.

The number of citations varies greatly between Scopus and Google Scholar. For the publications in the CLEF proceedings, Google Scholar finds almost nine times more citations than Scopus. When examining the distribution of citations over the years, Scopus indicates a variation in the number of citations, while Google Scholar shows a relative stability from 2005 onwards. For publications regarding ImageCLEF resources, Google Scholar finds almost five times more citations than Scopus. These peak for papers published in 2006 and 2004, mainly due to three publications that describe the creation of test collections that were used extensively in ImageCLEF in the following years, and thus attracted many citations. Overall, Google Scholar indicates that the total number of citations over all 249 publications in the considered dataset are 2,147, resulting in 8.62 average cites per paper. This is comparable to the findings of the study on the scholarly impact of TRECVID [Thornley et al., 2011], with the difference that they consider a much larger dataset of publications that also includes all TREC-derived papers.

Table 7: Overview of ImageCLEF publications 2004-2010 and their citations.

	Year	CLEF proceedings			ImageCLEF resources			All		
		papers	citations	h-index	papers	citations	h-index	papers	citations	h-index
S c o p u s	2004	5	13	2	4	31	3	9	44	4
	2005	20	50	4				20	50	4
	2006	25	24	3	3	28	1	28	52	3
	2007	27	25	2	6	29	2	33	54	3
	2008	29	18	3	5	22	2	34	40	3
	2009	45	14	2	2	4	1	47	18	2
	2010	44	38	4	11	7	2	55	45	4
	Total	195	182	6	31	121	5	226	303	9
G o o g l e	2004	5	65	3	5	105	4	10	170	6
	2005	20	210	8	5	47	4	25	257	10
	2006	25	247	7	8	144	5	33	391	9
	2007	27	259	7	10	76	4	37	335	9
	2008	29	249	7	7	73	5	36	322	9
	2009	45	284	7	7	53	4	52	337	9
	2010	44	259	7	12	76	6	56	335	10
	Total	195	1573	18	54	574	13	249	2147	22

Next, the impact of publications in the two domains studied in ImageCLEF, medical and general images, is investigated. Figure 4 compares the relative number of publications with the citation frequency for the domains. It should be noted that some publications examine

both domains at once, e.g., participants' papers presenting their approaches in ImageCLEF tasks that represent both domains, or overview papers reporting on all tasks in a year. Therefore, the sum of publications (citations) in Figure 4 is not equal to the total listed in Table 7. Overall, the publications in the medical domain appear to have a slightly higher impact. To gain further insights, Figure 5 drills down from the summary data into the time dimension. At first, publications relating to the general domain dominate, with those relating to the medical domain increasing as the corresponding tasks establish themselves in the middle of the time period, while more recently there is again a shift towards the general domain. Scopus indicates that the impact of ImageCLEF publications that are related to the medical domain is particularly significant between 2006 and 2008. This is mostly due to number of overview papers regarding the medical image annotation task published both in the CLEF proceedings and elsewhere, and also because Scopus does not index some of the ImageCLEF publications regarding general images that are found by Google Scholar. For Google Scholar, on the other hand, the distribution of citations appears to be mirroring that of the publications in the two domains.

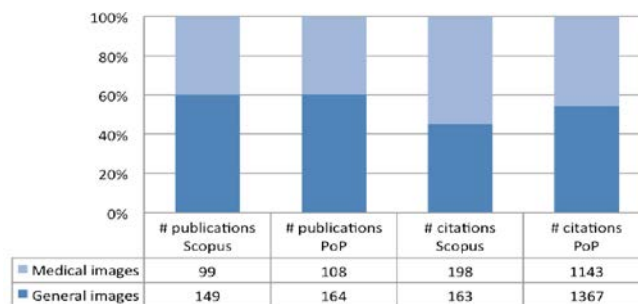


Figure 4: Relative impact of ImageCLEF publications in the two domains.

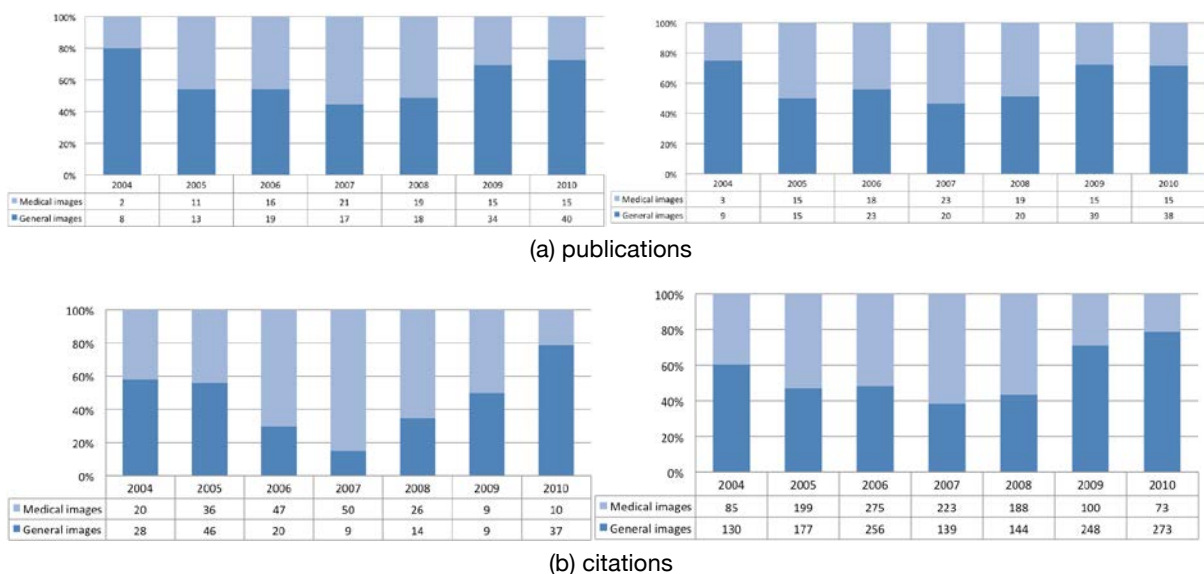


Figure 5: Relative impact of ImageCLEF publications in the two domains over the years.

Finally, Figure 6 depicts the distribution of citations for each of the ImageCLEF tasks (listed in Table 6) over the years. Similarly to above, a publication may cover more than one task.

For all tasks, there is a peak in their second or third year of operation, followed by a decline. The exception is the object and concept recognition task, which attracts significant interest in its fourth year when it is renamed as photo annotation task and employs a new collection consisting of Flickr images and new evaluation methodologies. These novel aspects of the task result not only in increased participation (see Table 6), but also strengthen its impact. Overall, the photographic retrieval, the medical image retrieval, and the medical image annotation tasks have had the greatest impact.

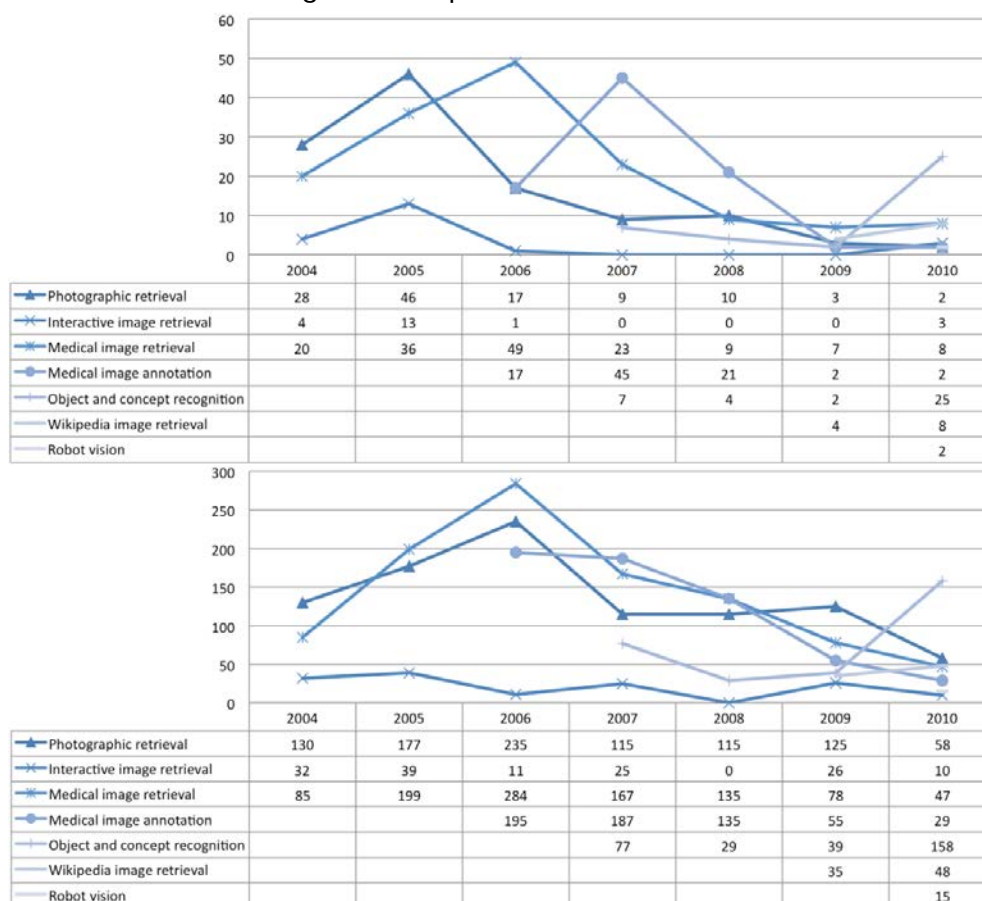


Figure 6: Citation trends per ImageCLEF task, Scopus (top) and PoP (bottom).

7.1.3 Conclusions

This work aims at analysing the scholarly impact of the ImageCLEF image retrieval evaluation campaign. Both Scopus and Google Scholar are used to obtain the number of papers published in the course of ImageCLEF and their citations. This preliminary analysis concentrates on the CLEF post-workshop proceedings, as the CLEF working notes are not indexed by Scopus, and therefore a fair comparison between Scopus and Google Scholar, one of the goals of this study, would not have been possible. A few additional papers written by the organisers about the main workshop outcomes are added. A total of 249 publications were analysed obtaining 2,147 citations in Google Scholar and 303 in Scopus. The analysis also shows that tasks usually take a year to attract a larger number of

participants but impact and participation usually drop after three years unless the task or the collection changes.

This preliminary analysis shows ImageCLEF's significant scholarly impact through the substantial numbers of its publications and their received citations. ImageCLEF data have been used by over 200 research groups, many techniques have been compared during its campaigns, while its influence through imposing a solid evaluation methodology and through use of its resources goes even further.

HES-SO in collaboration with the University of Padua are currently extending this work towards (i) automating the process as much as possible, (ii) including more ImageCLEF publications (the preliminary study only included a subset of all possible ImageCLEF-related publications), and (iii) assessing the scholarly impact of the whole CLEF evaluation campaign.

7.2 CLEF campaign website: web design qualitative analysis

The CLEF campaign website currently maintains information on the evaluation activities that have been conducted over the first ten years of CLEF (2000-2009). For each CLEF edition, the website maintains information on tracks, tasks in a track, contribution of the participants (notebook papers, presentations), and links to the publications concerning CLEF activities and to the workshop proceedings (which contain revised versions of the notebook papers). The website is intended both for CLEF participants and for researchers addressing the problem of the design, implementation, and evaluation of multimedia and multilingual retrieval systems. In order to adopt the CLEF website as a source of informative resources and relationships among resources:

- ✧ A qualitative analysis on the current version of the CLEF website was carried out to gain indications on the way heterogeneous types of users, i.e., users with different levels and fields of expertise, perceive the website. The outcomes of the analysis were used in redesigning the CLEF website to make it more responsive to user requirements.
- ✧ The website is being redesigned, thus supporting activities in an evaluation campaign, archiving its outcomes, and allowing its future integration with the PROMISE open evaluation infrastructure from which data and knowledge will be gathered.

This section reports on the qualitative analysis that was performed. The analysis was based on the website quality model proposed by Roberto Polillo and presented in a summarized version in [Polillo, 2005] and in a more detailed version in [Polillo, 2004]. The model concerns the *external quality* of the website, i.e., the quality perceived by the user.

This analysis is partly based on the assignments of students of the “Master Degree in Strategies on Communication”⁶ of the University of Padua, Italy, who attended the course on “Website Design” taught during the first semester of the 2010-2011 academic year by

⁶ A Master Degree course in Italian is named “Laurea Magistrale”; the course in Strategies on Communication is named “Laurea Magistrale in Strategie della Comunicazione” (LMSDC), information on this master degree course is available at the URL: http://www lettere.unipd.it/magistrali/lmsgs/mag_lmsgs.html

Maristella Agosti⁷. One part of the course on Website Design addresses the different aspects to be taken into account when evaluating a website. During this part of the course Polillo's model for evaluating the quality of a website was introduced and then adopted by the students to carry out the analysis on a set of assigned websites.

Section 7.2.1 briefly reviews Polillo's model and the methodology adopted for gathering indications from the students, while Section 7.2.2 reports on the outcomes of the qualitative analysis and possible actions for the website redesign.

7.2.1 Website quality model and quality analysis methodology

The website quality model proposed in [Polillo, 2004; Polillo, 2005] and exploited for the analysis reported in this section is based on seven main characteristics that are evaluated on a scale of 0-4. The following table briefly introduces the seven main characteristics of the website quality model adopted.

Table 8: main characteristics taken into account for the analysis

Characteristic	Sub-characteristic	Main questions that need to be answered when conducting the evaluation
Architecture	Structure	Is the informative structure of the site adequate?
	Site map	Does a site map exist that clearly represents the site structure?
	Navigation	Is the navigation of the site adequate?
Communication	Home page	Is the homepage clearly communicating the site objectives?
	Brand image	Is the site coherent with the brand image?
	Graphics	Are the site graphics adequate?
Functions	Adequacy	Are the site functions adequate?
	Correctness	Are the site functions correct?
Content	Categorization/labelling	Is the information classified in an adequate way?
	Style	Is the text style adequate for a web presentation?
	Information	Is the information adequate, pertinent, reliable and up to date?
	Localization	Is the site correctly localized?
Management	Availability	Is the site always up and available?
	Monitoring	Is use of the site monitored?
	Update	Is the site constantly updated and improved?
	Relationships with the users	Are the relationships with the users adequately

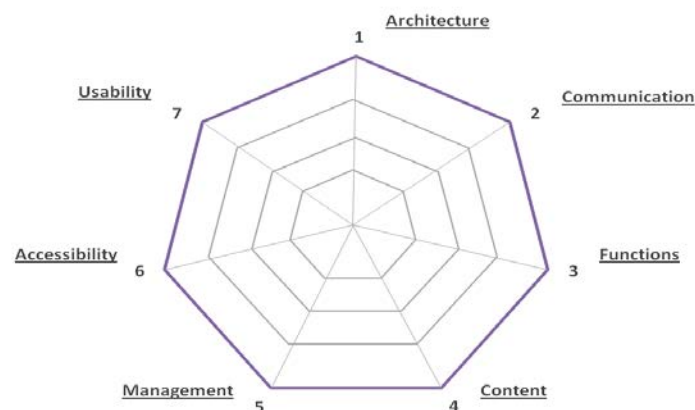
⁷

Information on the course is available at the URL:

http://www.lettere.unipd.it/infolettere/pub/docente_public.php?doc=174

Characteristic	Sub-characteristic	Main questions that need to be answered when conducting the evaluation
		presided?
Accessibility	Access time	Is the access time adequate?
	Possibility of retrieval	Is it easy to find the site?
	Browser independence	Is the site accessible through different browsers?
	Accessibility for impaired people	Is the site accessible for impaired people?
Usability	Efficacy	Can the user reach his objectives in an accurate and complete way?
	Efficiency	Is the level of user effort acceptable?
	User satisfaction	Is the site comfortable and acceptable for the user?

When the analysis of a website is completed and the table is filled in, a summary of the results can be represented in a graphical way using a “radar” type of representation of the seven characteristics. An example of this type of graphical representation, also named *radar*



diagram, is shown below.

Figure 8: radar diagram.

The scale of evaluation of each characteristic is from 0 to 4, where each subinterval of the figures represents the following:

- ⤴ 3-4: excellent
- ⤴ 2-3: good
- ⤴ 1-2: satisfactory
- ⤴ 0-1: unsatisfactory.

A radar diagram at a higher level of detail can be also adopted, specifically based on scores on all the sub-characteristics reported in the table above; in the remainder of this section we

will adopt the radar diagram based on the seven main characteristics because data on the sub-characteristics are not available for all the students who performed the analysis.

The students were required to study the model and learn to apply and use it. Therefore, the model was taught and presented over a number of lectures, where several websites were presented and initially analysed. Afterwards, a number of relevant websites were presented during laboratory lectures; following those lectures the students were required to analyse those relevant websites for a period of two weeks and afterwards prepare a written assignment reporting their findings.

One of the websites considered of relevance during the course of the 2010-2011 academic year was the CLEF campaign website. The students were presented with the main objectives of the site during laboratory lectures. During these lectures, other websites were analyzed. After those lectures, students were asked to prepare a written assignment. The target of the assignment was to analyze five relevant websites and report the findings, with one of the websites being the CLEF campaign website. The report was prepared using the adopted website quality model. In this way the students reported their findings all using the same model; this makes their findings interesting in their own right but also comparable to each other.

7.2.2 CLEF campaign website qualitative analysis

This section reports on results of the analysis of the website. The number of assignments that were analysed was 54. The assignments of 39 among the 54 students were selected, since the students in this subset provided a weighted score for each of the seven characteristics of Polillo's model. The following table reports the mean score per characteristic computed over all the 39 assignments. The score of a characteristic was computed as the weighted sum of the sub-characteristic scores, where the weight was assigned to each sub-characteristic by the students.

Table 9: scores of the CLEF web page.

Characteristic	Mean Score
Architecture	2.13
Communication	2.04
Functions	2.67
Content	2.76
Management	2.81
Accessibility	2.51
Usability	2.44

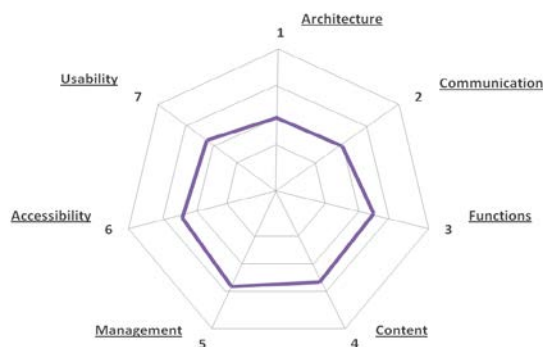


Figure 9: radar diagram for the CLEF web page

The radar diagram in Figure 9 provides a graphical representation of the mean values obtained for the seven characteristics.

The values reported in the Table 9 show that the quality of the website is on average perceived as good. The content is perceived as reliable by the students, even if it is not in their domain of expertise. The management obtained a good value, which is mainly due to the frequency of updates and the availability of the website; the quality of the management, can be further increased by improving the monitoring activities, e.g. fixing broken links or removing obsolete information. The website is perceived by the students as an archive of information and resources related to the evaluation activity; an expert user familiar with the workflow of an evaluation campaign can successfully reach the information he is looking for, even if the navigation paths that link diverse informative resources can be quite structured and the user effort required can be high. These are some of the motivations for the low values obtained for architecture and communication, i.e., characteristics that concern the domain of expertise of the students. The students' remarks suggest a redesign of the informative structure to make relationships among informative resources explicit, which is actually the mentioned PROMISE objective and one of the motivations for the analysis reported in this section.

Specific detailed remarks for each of the seven characteristics and their sub-characteristics are provided in Appendix V: Qualitative analysis of CLEF campaign website.

8 Outlook on future evaluation activities: CLEF 2011

This section provides an outlook on the future evaluation activities in the second year of PROMISE by outlining the steps taken towards the organization of the CLEF 2011 Conference and its current status, by briefly describing the selection process for the CLEF 2011 Labs, and by listing the selected labs. This section provides only a brief summary of these activities; further details will be provided in Deliverable 7.5 “Second PROMISE Annual Conference and Proceedings”.

CLEF 2011 conference: The CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation – an event organized by PROMISE – will take place in Amsterdam, The Netherlands on September 19-22, 2011. In summary:

1. There were 23 submissions (19 full papers and 4 short papers), a slight increase with respect to last year. In total, 14 papers (10 full and 4 short) were accepted with an overall acceptance rate of 60,87%.
2. There will be two keynote talks: Elaine Toms, University of Sheffield, and Omar Alonso, Microsoft.
3. There will be a “community” session aimed at offering insights in funding, networking, and infrastructure opportunities.
4. The accepted papers and the abstracts of the invited talks will be published by Springer in their Lecture Notes for Computer Science (LNCS) series. The Conference Proceedings will be ready in time for the conference in September and distributed to conference attendees.
5. When registering for CLEF 2011, an account is created for you in a social network portal dedicated to the conference. This allows participants to communicate with each other in an easy way before the conference starts.

CLEF 2011 labs: Interested lab organizers can propose one of two kinds of lab: a “campaign-style” track for evaluating a certain IR task or a workshop that considers other evaluation issues and which follows a more exploratory pattern. The selection committee was bound by the following instructions: to limit the number of tasks per benchmarking activity; to consider re-positioning proposed benchmarking activities that have important “gaps” (in task definition, feasibility, availability of data, relevant expertise or the organizers, etc.) as half-day workshops; to avoid overlap with related benchmarking activities at sister-events such as TREC, NTCIR and INEX; to re-adjust the length of a benchmarking activity or workshop based on relevance and expected interest; to limit the total length of all activities to 2 full days; to limit the total number of parallel lab events (i.e., benchmarking activities and/or workshops) at the CLEF 2011 conference to 3.

1. There were 9 lab proposals: 6 were accepted as “campaign-style” or benchmarking labs and 1 was accepted as an exploration workshop, resulting in an acceptance rate of 77% (=7/9).
2. The CLEF 2011 Labs are the following:
 - i. **ImageCLEF** – Cross-Language Image Retrieval: evaluation of retrieval from visual collections; both textual and visual retrieval techniques are exploitable.

Four challenging tasks are foreseen: 1) retrieval from a collection of Wikipedia images with textual annotations and topics in several languages; 2) medical image retrieval with visual, semantic and mixed topics in several languages with a data collection from the scientific literature; 3) visual classification of leaf images for the identification of plant species; 4) a photo annotation task that investigates automated semantic annotation based on visual information with approaches based on Flickr user tags and multimodal approaches.

- ii. **QA4MRE** - Question Answering for Machine Reading Evaluation: evaluation of Machine Reading abilities through Question Answering and Reading Comprehension Tests.
 - iii. **MusiCLEF**: the goal is to promote the development of novel methodologies for music access and retrieval that combine content-based information, automatically extracted from music files, with contextual information, provided by users through tags, comments, reviews, possibly in different languages.
 - iv. **LogCLEF**: the goal is the analysis and classification of queries in order to understand search behavior in multilingual contexts; it consists of three tasks: 1) language identification, 2) query classification, and 3) Success of a query.
 - v. **PAN** – Uncovering Plagiarism, Authorship, and Wikipedia Vandalism: it consists of three tasks: 1) Plagiarism Detection, 2) Author Identification, and 3) Wikipedia Vandalism Detection.
 - vi. **CLEF-IP** - IR in the IP domain: using a collection of more than 2 million patent documents in XML format with content in English, German, and French that also include patent images, five tasks are foreseen 1) the Prior Art Candidate Search task, 2) the Image-based Document Retrieval task, 3) The Classification task, 4) the Refined Classification task, and 5) the Image-based Classification task.
 - vii. **CHiC 2011** – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage: this workshop aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems by surveying all the evaluation efforts in the cultural heritage field as well as defining user scenarios and identifying possible relevant metrics.
3. More than 150 groups have shown interest in the CLEF 2011 labs, distributed as follows: ImageCLEF 73, PAN 45, CLEF-IP 25, QA4MRE 24, LogCLEF 17, MusiCLEF 4, and CHiC Workshop 3.
 4. ImageCLEF is the most popular lab able to attract many participants not only from Europe but from the entire world.
 5. Overall, although most of the participants are still coming from Europe, an increasing interest in CLEF all around the world can be registered.
 6. All working notes for all labs will be published online for the conference.

9 References

- [Amigó et al., 2010] Enrique Amigó, Javier Artiles, Julio Gonzalo, Damiano Spina, Bing Liu, Adolfo Corujo: WePS3 Evaluation Campaign: Overview of the Online Reputation Management Task. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Artiles et al., 2010] Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, Enrique Amigó: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Braschler et al., 2010] Martin Braschler, Donna Harman: Introduction to the CLEF 2010 labs. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Cleverdon, 1959] Cyril W. Cleverdon. The evaluation of systems used in information retrieval. In Proceedings of the International Conference on Scientific Information (Vol. 1), pages 687–698. National Academy of Sciences, National Research Council, 1959.
- [Mandl et al., 2010] Thomas Mandl, Giorgio Maria Di Nunzio, Julia Maria Schulz: LogCLEF 2010: the CLEF 2010 Multilingual Logfile Analysis Track Overview. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Müller et al., 2010] Henning Müller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Joe Reisetter, Charles E. Kahn Jr., William R. Hersh: Overview of the CLEF 2010 Medical Image Retrieval Track. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Nowak & Huiskes, 2010] Stefanie Nowak, Mark J. Huiskes: New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.

- [Peñas et al., 2010] Anselmo Peñas, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, Corina Forascu, Cristina Mota: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Piroi, 2010a] Florina Piroi: CLEF-IP 2010: Prior Art Candidates Search Evaluation Summary, Technical Report IRF-TR-2010-0003, July 2010.
- [Piroi, 2010b] Florina Piroi: CLEF-IP 2010: Classification Task Evaluation Summary Technical Report IRF-TR-2010-00004, August 2010.
- [Piroi, 2010c] Florina Piroi: CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Polillo, 2004] Roberto Polillo. Il check-up dei siti web – valutare la qualità per migliorarla. Apogeo, Milano, 2004.
- [Polillo, 2005] Roberto Polillo. Un modello di qualità per i siti web. Mondo digitale, n.2, giugno 2005, pp. 32-44. Available online at the URL:
http://www.mondodigitale.net/Rivista/05_numero_tre/Polillo_p._32-44.pdf
- [Popescu et al., 2010] Adrian Popescu, Theodora Tsikrika, Jana Kludas: Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Potthast et al., 2010a] Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, Paolo Rosso: Overview of the 2nd International Competition on Plagiarism Detection. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Potthast et al., 2010b] Martin Potthast, Benno Stein, Teresa Holfeld: Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.

- [Pronobis et al., 2010] Andrzej Pronobis, Marco Fornoni, Henrik I. Christensen, Barbara Caputo: The Robot Vision Track at ImageCLEF 2010. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Rowe et al. 2010] Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. Technical Report Project Number 0211875, RTI International, 2010.
- [Sorg et al., 2010] Philipp Sorg, Philipp Cimiano, Antje Schultz, Sergej Sizov: Overview of the Cross-lingual Expert Search (CriES) Pilot Challenge. In Martin Braschler, Donna Harman, Emanuele Pianta (Eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy, 2010.
- [Thornley et al., 2011] Claire V. Thornley, Andrea C. Johnson, Alan F. Smeaton, Hyowon Lee: The scholarly impact of TRECVID (2003-2009). JASIST, 62(4):613{627, 2011.
- [Tsikrika et al., 2011] Theodora Tsikrika, Alba G. Seco de Herrera, and Henning Müller. Assessing the Scholarly Impact of ImageCLEF. In Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2011), 19-22 September, Amsterdam, The Netherlands, 2011.

Appendix I: Questionnaires sent to CLEF 2010 Labs organizers

1. CLEF 2010 Labs

https://docs.google.com/spreadsheet/viewform?pli=1&hl=en_US&formkey=dHZtanhGNzJPQTZhRWgxRzhRTzU5Y0E6MQ#gid=0

2. CLEF 2010 Labs: collections

https://docs.google.com/spreadsheet/viewform?hl=en_US&formkey=dFBsWmZEOWJkQ0J1eGVnUHpXWG1fb3c6MA#gid=0

CLEF 2010 Labs

Please provide some information regarding the task/Lab you organised in 2010.

If one of the required questions does not apply to you, please enter "Not applicable" as an answer.

Thank you very much for you input!

*** Required**

Lab: *

Task: *

(if there are no tasks in the Lab, please enter the name of the Lab)

Main task organiser(s) (name, email, and affiliation): *

Person who filled in this form (name, email, and affiliation): *

How many years has this task been running at CLEF? *

(if it is the first year, please enter 1; if it is the second year, please enter 2; ...)

Has this task previously run as part of another evaluation campaign?

- ☐ Yes
☐ No

If the answer to the previous question is "Yes", please specify which evaluation campaign:

Main objective of the task? *

How would you best describe the task? *

- ☐ Retrieval
☐ Classification
☐ Question Answering
☐ Expert Search
☐ Other:

What is the retrieval unit for this task? *

i.e., what is the system supposed to retrieve, classify, find?

What constitutes a topic for this task? *

(briefly describe what a topic consists of)

Number of topics: *

Are the topics multilingual? *

- ☐ Yes
☐ No

If the answer to the previous question is "Yes", please specify which languages:

Number of results per topic: *

(max number of results that can be submitted per topic)

Number of registrations: *

Number of participations: *

(number of research groups that officially submitted results)

Number of return participations: *

(number of research groups that also participated in 2009)

Number of submissions allowed per participant: *

(max number of runs each participant is allowed to submit)

Number of submissions: *

(number of submitted runs in total)

Submission system: *

- ☐ submissions sent by email
- ☐ submissions uploaded to server
- ☐ DIRECT
- ☐ ImageCLEF submission system
- ☐ Other:

Ground truth generation: How many documents were assessed? *

(e.g., "All documents in the collection", "Pooled documents; pools created with depth of 50 documents per run", ...)

Ground truth generation: How many assessors were employed? *

(if the assessments were generated automatically, please write "automatic relevance assessments" in all questions regarding ground truth generation.)

Ground truth generation: Who were the assessors? *

(Please provide a brief description who the assessors were; e.g., "students", "domain experts", "crowdsourcing", ...)

Ground truth generation: How much time did the assessors spend? *

(Please provide an estimate for the time it took to create the ground truth.)

Did you (the organisers) provide a baseline? *

- ☐ Yes
☐ No

Were there any tools offered to the participants? If yes, which ones? *

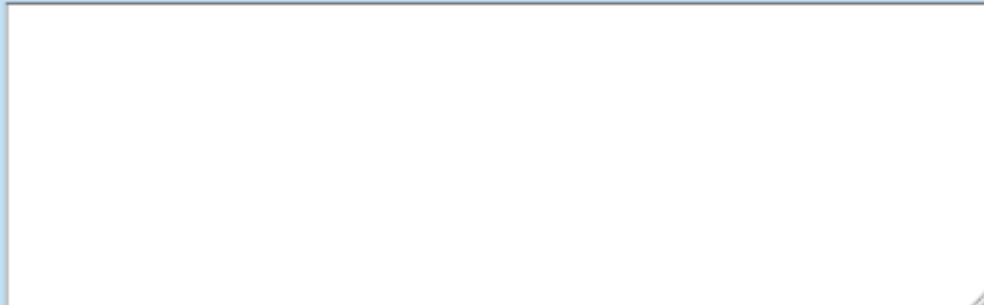
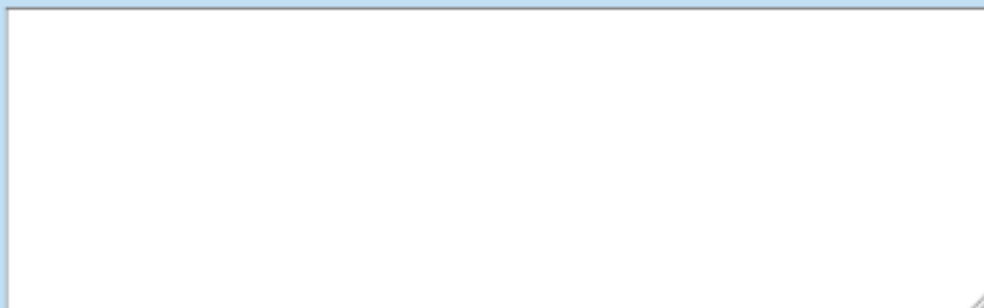
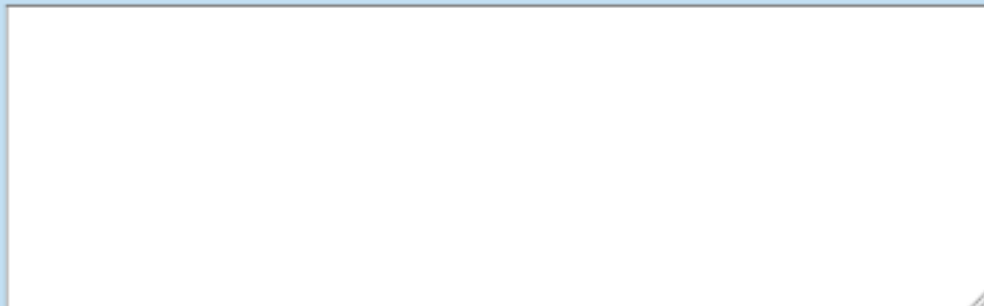
How many participants used these tools? *

How were the evaluation measures computed? *

- ☐ trec_eval
☐ in-house implementation
☐ DIRECT
☐ Other:

Main differences/advancements from 2009: *

(e.g., collection size and characteristics, number of topics and their characteristics, new objectives, ...)

A large, empty rectangular box with a thin black border, intended for text input. It is located below the first section header.**Main problems (from an organisational point of view): ***A large, empty rectangular box with a thin black border, intended for text input. It is located below the second section header.**Main trends (among the participants' approaches): ***A large, empty rectangular box with a thin black border, intended for text input. It is located below the third section header.

Main experimental outcomes (based on the participants' approaches): *

Anything else?

Submit

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

CLEF 2010 Labs: collections

Please provide information on the collection used in your task/Lab.

If more than one collections were used, please fill in a separate form for each one.

If one of the required questions does not apply to you, please enter "Not applicable" as an answer.

Thank you very much for you input!

*** Required**

Lab: *

Task(s): *

(if there are no tasks in the Lab, please enter the name of the Lab)

Collection name: *

Briefly describe the collection. *

(e.g., What does it contain? textual documents, web pages, wikipedia pages, images, tweets, search log records or other? How was it obtained? Did you crawl it yourselves? Please add anything you consider important.)

Number of documents in collection: *

Size of collection:

(actual size of the collection in MB, GB,)

Is the collection multilingual? *

- ☐ Yes
☐ No

If the answer to the previous question is "Yes", please specify which languages:

Was the collection created for this task? *

- ☐ Yes
☐ No

How many years has the collection been used in this task? *

(if it was used for the first time in 2010, please enter 1; ... This question refers to collection used in 2010 as a whole. If a subset of the collection was used in previous years of this task, please fill in the next question)

Have parts of the collection used in 2010 been used in previous years of this task?

- ☐ Yes
☐ No

If the answer to the previous question is "Yes", please provide some details on which parts:

(e.g., "The 2010 collection consists of 200,000 images, 100,000 of which were part of the 2009 collection.")

Has this collection been used in other tasks/Labs? *

- ☐ Yes
☐ No

If the answer to the previous question is "Yes", please specify which task/Lab:

Submit

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Appendix II: Participation in the CLEF 2010 labs

Table 8: Participation to the CLEF 2010 labs

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
CLEF-IP	Patent Classification	1	12	7	not applicable	8	27	submissions uploaded to a server
CLEF-IP	Prior Art Candidates Search	2	22	9	7	8	25	submissions uploaded to a server
ImageCLEF	Medical image retrieval	7	51	16	8	10 per category	155	ImageCLEF submission system
ImageCLEF	Photo Annotation	5 ⁸	54	17	10	5	63	ImageCLEF submission system
ImageCLEF	Robot Vision	2	43	7	3	unrestricted	55	ImageCLEF submission system
ImageCLEF	Wikipedia image retrieval	3	50	13	2	20	127	ImageCLEF submission system
PAN	Plagiarism	1	44	18	5 ⁹	unrestricted	27	RapidShare

⁸ The first annotation task was organized in 2006. However, the collections and layout of the task significantly changed in 2009 from a pure visual task to a multi-modal task.

⁹ This is the first time the Plagiarism Detection task is organized under the auspices of CLEF, but overall it is the second time such a task is organized.

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Detection							
PAN	Wikipedia Vandalism Detection	1	15	9	0	unrestricted	15	RapidShare
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	2 of ResPubliQA; 7 of QA@CLEF	24	13	8	2	49	CELCT submission system
WePS	Online reputation management	1	5	5	0 ¹⁰	unrestricted	16	submissions sent by email
WePS	Clustering Attribute Extraction	1	34	8	3 ¹²	5	27	submissions sent by email
CriES	CriES Pilot Challenge	1	4	4	not applicable	3	9	submissions sent by email
LogCLEF	LogCLEF	2	15	7	3	not applicable	not applicable	results presented at the Lab

¹⁰

WePS ran for the first time in CLEF in 2010, but had previously run twice as a workshop: WePS-1 (at Semevel) and WePS-2 (at the WWW conference).

Appendix III: Main outcomes of the CLEF 2010 Labs

Table 9: Main advancements in the CLEF 2010 Labs

Lab	Task	Task type	Main differences/advancements from 2009
CLEF-IP	Patent Classification	Classification	Task first ran in 2010
CLEF-IP	Prior Art Candidates Search	Retrieval	<ul style="list-style-type: none"> larger collection fewer topics each topic is now one document (the patent application document) and not a composed document out of various patent documents; this makes the task more realistic
ImageCLEF	Medical image retrieval	Retrieval	<ul style="list-style-type: none"> larger collection more case-based topics
ImageCLEF	Photo Annotation	Classification	<ul style="list-style-type: none"> larger number of concepts inclusion of Flickr user tags differentiation of runs into textual, visual and multi-modal runs new objectives <ul style="list-style-type: none"> Do multi-modal approaches outperform text only or visual only approaches? Which approaches are best for which kind of concepts?
ImageCLEF	Robot Vision	Classification	<ul style="list-style-type: none"> new collection image sequences acquired by a stereo camera, as opposed to a perspective camera in 2009 larger number of areas to be recognised
ImageCLEF	Wikipedia image retrieval	Retrieval	<ul style="list-style-type: none"> new larger collection (150,000 images in 2009 vs. 237,000 images in 2010) multilingual collection (the 2009 collection was monolingual) collection includes not only the user-generated annotations, but also the

Lab	Task	Task type	Main differences/advancements from 2009
			<p>Wikipedia articles in which the images were embedded.</p> <ul style="list-style-type: none"> • topics also became multilingual and more sample images were included. • Investigation of not only multimodal, but also multilingual approaches
PAN	Plagiarism Detection	Retrieval	<ul style="list-style-type: none"> • error corrections in the 2009 corpus
PAN	Wikipedia Vandalism Detection	Classification	Task first ran in 2010
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	Question Answering	<ul style="list-style-type: none"> • addition of a portion of the EUROPARL collection • participants had the opportunity to return both paragraph and exact answers as system output.
WePS	Online reputation management	Document filtering	Task first ran in 2010
WePS	Clustering Attribute Extraction	Document Clustering Information Extraction	<ul style="list-style-type: none"> • larger and more diverse test bed • the Attribute Extraction task now requests to relate each attribute to a person (cluster of documents) instead of just listing the attributes obtained from each document
CriES	CriES Pilot Challenge	Expert Search	Task first ran in 2010
LogCLEF	LogCLEF	Log Analysis	<ul style="list-style-type: none"> • larger collection • identification and definition of common problems and tasks

Table 10: Main trends in the approaches employed by the participants to the CLEF 2010 Labs and the main experimental outcomes.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
CLEF-IP	Patent Classification	<ul style="list-style-type: none"> Use of SVM and k-NN classifiers. 	<ul style="list-style-type: none"> Measure values were very good, scores being very high. Systems that did classification using a retrieval system did not perform as well as systems that used exclusively classification engines.
CLEF-IP	Prior Art Candidates Search	<ul style="list-style-type: none"> Use of tf-idf and text analysis tools, experiment with searching within various parts of the documents. Only few participants used the information-carrying metadata in the collection's documents. Where multilinguality was involved, participants chose to use Google Translate for query translation. 	<ul style="list-style-type: none"> The usual text-analysis/retrieval tools only are not good enough for this specific retrieval domain. The best results were given by a system that added more patent specific information processing to the typical text-analysis tools.
ImageCLEF	Medical image retrieval	<ul style="list-style-type: none"> Mapping of text onto medical ontologies such as MeSH or UMLS. Visual retrieval with salient point-based features; this obtains better results than other global features. 	<ul style="list-style-type: none"> Text is much better than visual retrieval. Visual retrieval sometimes obtains good early precision. Filtering based on (even erroneous) visual classification improves results.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
ImageCLEF	Photo Annotation	<ul style="list-style-type: none"> In the visual and multimodal configurations, discriminative classifiers and nearest-neighbour approaches are dominating. Most teams applied SIFT and color SIFT features for visual classification. Some teams additionally used global features like color histograms. Many teams analyzed the X most occurring Flickr tags and built a binary vector of the most occurring tags for each image to incorporate textual information. 	<ul style="list-style-type: none"> The multimodal approaches got the best scores for 61 out of 93 concepts, followed by 30 concepts that could be detected best with the visual approach and two that won with a textual approach. The multimodal approaches outperformed visual and textual configurations for all teams that submitted results for more than one configuration
ImageCLEF	Robot Vision	Not available.	Not available.
ImageCLEF	Wikipedia image retrieval	<ul style="list-style-type: none"> More multimodal and multilingual approaches are being applied. Use of external sources to enhance retrieval (e.g., DBpedia, Flickr etc.). 	<ul style="list-style-type: none"> Multilingual approaches are more successful than monolingual ones. The effectiveness of multimodal approaches varies: some judge the visual features as very helpful, whereas others find that they are helping only a little. This could be due to using the visual features as an input of the same

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
			importance as the textual features does not work not well, whereas boosting, reranking and query expansion with the help of visual features achieves good results.
PAN	Plagiarism Detection	<ul style="list-style-type: none"> Participants did not apply keyword retrieval but compared all pairs of documents exhaustively, which is not practical 	<ul style="list-style-type: none"> Artificially generated plagiarism can be found with high performance. Manually generated plagiarism is a lot harder to find.
PAN	Wikipedia Vandalism Detection	<ul style="list-style-type: none"> Various different paradigms for features have been employed, some content-based, some context-based. No participant employed two paradigms at the same time. 	<ul style="list-style-type: none"> Solving vandalism detection is within reach, but a combination of all feature paradigms is necessary.
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	Not available.	Not available.
WePS	Online reputation management	<ul style="list-style-type: none"> Learning language models from the company's URL and additional resources. Topic classification in order to infer the 	<ul style="list-style-type: none"> Considering additional sources like Google results or WordNet seems to be useful. Linguistic aspects of the company mention are also very indicative.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
		amount of relevant documents.	<ul style="list-style-type: none"> It is possible to define a general approach to estimate approximately the presence of a company name in Twitter. Bootstrapping methods seems to be useful, specially for highly ambiguous company names.
WePS	Clustering Attribute Extraction	<ul style="list-style-type: none"> Many systems include Hierarchical Agglomerative Clustering (HAC) as part of their system pipeline. One team intentionally departs from the usage of HAC and experiments with the k-Medoids clustering method. Another team compared three clustering methods (Lingo, HAC, and 2 steps HAC) using basic features extracted from the web pages. 	<ul style="list-style-type: none"> The best scoring system obtains balanced results in both precision and recall, while the rest of the participants have biased scores towards one or other metric. The Unanimous Improvement Ratio results confirmed that only the top two systems in the ranking make a robust improvements (independent of the weighting of Precision and Recall). As in the previous WePS campaigns, the correct selection of a cluster stopping criterion is a key factor in the performance of systems. Unlike previous WePS campaigns, almost all the systems obtained scores above the baselines.
CriES	CriES Pilot Challenge	<ul style="list-style-type: none"> Application of standard retrieval approaches to the novel setting. 	<ul style="list-style-type: none"> Combinations of retrieval models that exploit the social graph are more successful than approaches that only rely on text.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
LogCLEF	LogCLEF	<ul style="list-style-type: none"> Every participant presented different approaches to log analysis. 	<ul style="list-style-type: none"> Presented approaches addressed: language identification, named entity recognition in queries, query classification, and definition of success of a search. A basic set of tools and frameworks shared within the community is clearly needed so that individual researchers do not have to re-invent the wheel.

Appendix IV: CLEF 2010 Labs Test Collections

List of collections in the CLEF 2010 labs:

- ✧ **CLEF-IP 2010 collection:** A collection of patent documents from the EPO (European Patent Office), published up until 2001. The collection contains only textual documents and was obtained from a third party provider.
- ✧ **RSNA images:** A collection of medical images obtained from Journals of the Radiological Society of North America, namely Radiographics and Radiology.
- ✧ **ImageCLEF VCDT 2010 collection:** A collection of 18,000 images from the MIR Flickr collection including EXIF data and Flickr user tags.
- ✧ **COLD-Stockholm:** A collection of image sequences acquired using a robot platform equipped with a stereo camera system. The acquisition was performed on an office environment, consisting of 36 areas belonging to 12 different semantic and functional categories. The robot was manually driven through the environment and each data sample was then labelled as belonging to one of the areas according to the position of the robot during acquisition (rather than contents of the images).
- ✧ **ImageCLEF 2010 Wikipedia collection:** A collection of 237,434 Wikipedia images, their user-provided annotations and the Wikipedia articles that contain these images. The collection was built to cover similar topics in English, German and French and it is based on the September 2009 Wikipedia dumps. Images are annotated in none, one or several languages and, wherever possible, the annotation language is given in the metadata file. The articles in which these images appear were extracted from the Wikipedia dumps and are provided as such.
- ✧ **PAN Plagiarism Corpus 2010 (PAN-PC-10):** The corpus contains books downloaded from the Project Gutenberg in the form of text documents.
- ✧ **PAN Wikipedia Vandalism Corpus 2010 (PAN-WVC-10):** The collection contains meta information about edits on Wikipedia articles. Moreover, it contains the edited articles as wikitext documents.
- ✧ **ResPubliQA 2010:** The collection contains two multilingual parallel corpora: a subset of the JRC-ACQUIS Multilingual Parallel Corpus and a small portion of the EUROPARL collection. JRC-ACQUIS is a freely available parallel corpus containing the total body of European Union (EU) documents, mostly of legal nature. It comprises contents, principles and political objectives of the EU treaties; the EU legislation; declarations and resolutions; international agreements; and acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more. EUROPARL is a collection of the Proceedings of the European Parliament dating back to 1996 and comprises parallel texts in each of the official languages of the European Union. The collection was obtained by crawling the European Parliament site selecting documents from January 2009 onwards.
- ✧ **WePS-3 ORM:** Each organization in the dataset is associated with the company name and its homepage. The input information per tweet consists of a tuple

containing: the tweet identifier, the entity name, the query used to retrieve the tweet, the author identifier and the tweet content. The training and test corpora contain 100 different company names.

- ✧ **WePS-3:** The dataset consists of the top 200 web search results from the Yahoo! API for 300 different ambiguous person names. These web results were downloaded and archived with their corresponding search metadata (search snippet, title, URL and position in the results ranking). The dataset also contains human assessments of the correct way to group these documents according to the different people mentioned with the same name. The names were obtained randomly from the US Census, Wikipedia and computer science conference program committees. In addition to that, names were included for which at least one person has one of the following occupations: attorney, corporate executive or realtor.
- ✧ **CriES Yahoo! Answers Collection:** Crawl of Yahoo! Answers containing questions and answers from the social community platform. The used dataset is a subset of an official crawl released by Yahoo!.
- ✧ **The European Library (TEL) logs:** From January 2007 until June 2008: about 1,870,000 records of action logs. From January 2009 until December 2009: about 2,600,000 records of action logs.
- ✧ **Deutscher Bildungsserver (DBS) logs:** The "Deutscher Bildungsserver" is a quality controlled internet directory for educational resources. A raw server log representing three months of activities on the portal is made available.

Table 11: Collections used in the tasks of the CLEF 2010 Labs.

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab	Parts of the collection used in previous years of the lab
CLEF-IP	Prior Art Candidate Search Patent Classification	CLEF-IP 2010	2,600,000	19 GB	EN, DE, FR	Yes	2	Patent documents published prior to 2000 were in the CLEF-IP 2009 collection.
ImageCLEF	Medical image retrieval	RSNA images	74,902	16 GB	EN	Yes	3	Collection first created in 2008. Each year a few thousand new images are added.
ImageCLEF	Photo Annotation	ImageCLEF VCDT 2010	18,000		EN ¹¹	Yes	2	Same image collection, but a smaller set of annotations was used in 2009.
ImageCLEF	Robot vision	COLD-Stockholm	9,592	4.5 GB	No text	Yes	1	No
ImageCLEF	Wikipedia image retrieval	ImageCLEF 2010 Wikipedia collection	237,434	25 GB	EN, DE, FR	Yes	1	No
PAN	Plagiarism	PAN-PC-10	27,075	5 GB	EN, DE, ES	Yes	1	No

¹¹ Some Flickr image user tags are in languages other than English.

Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab	Parts of the collection used in previous years of the lab
	Detection							
PAN	Wikipedia Vandalism Detection	PAN-WVC-10	32,439	1.4 GB	EN	Yes	1	No
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	ResPubliQA 2010	10,700	670 MB	EN, DE, FR, IT, PT, RO, ES, BG, NL	Yes	2	The JRC-ACQUIS subset was also used in 2009.
WePS	Online Reputation Management	WePS-3 ORM	20,000		EN	Yes	1	No
WePS	Clustering Attribute Extraction	WePS-3	60,000	520MB	EN	Yes	3	Yes
CriES	CriES Pilot Challenge	CriES Yahoo! Answers Collection	780,193	~600MB	EN, DE, FR, ES	Yes	1	No
LogCLEF	LogCLEF	The European Library (TEL) logs	~4,000,000 action log records	~2 GB	Queries in logs in any language, usually European.	Yes	2	A subset of ~1,870,000 action log records was used in 2009.
LogCLEF	LogCLEF	Deutscher Bildungsserver	~5 GB	~5 GB	(mostly) DE, (some) EN	Yes	1	No



PROMISE

Participative Research labOratory for Multimedia and
Multilingual Information Systems Evaluation



Lab	Task(s)	Collection	Number of documents	Size	Languages	Collection created for the lab	Number of years collection used in lab	Parts of the collection used in previous years of the lab
		(DBS) logs						

Table 12: Topics used in the tasks of the CLEF 2010 Labs.

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
CLEF-IP	Patent Classification	Classification	A patent application document, as it is filed in at a patent office, in XML to be classified according to the International Patent Classification System.	2,000 docs 639 classes	EN, DE, FR ¹²
CLEF-IP	Prior Art Candidates Search	Retrieval	A patent application document, as it is filed in at a patent office, in XML.	2,000	EN, DE, FR ¹⁴
ImageCLEF	Medical image retrieval	Retrieval	A multimedia query that consists of a textual part, the query title in three languages, and a visual part, one or several example images.	30	EN, DE, FR
ImageCLEF	Photo Annotation	Classification	Flickr images classified into concepts. A concept is an English term about a depicted entity, representational characteristics or affective terms in images, e.g. car, city, close-up, portrait, boring.	10,000 images 93 concepts	EN
ImageCLEF	Robot Vision	Classification	Images classified into semantic categories representing rooms and functional areas.	2,471 images	Not applicable

¹²

There are documents with content mostly in French, mostly in German, or mostly in English. Some XML elements are given only in English.

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
				9 categories	
ImageCLEF	Wikipedia image retrieval	Retrieval	A multimedia query that consists of a textual part, the query title in three languages, and a visual part, one or several example images.	70	EN, DE, FR
PAN	Plagiarism Detection	Retrieval	The document to be analyzed for plagiarism.	27,075	EN, DE, ES
PAN	Wikipedia Vandalism Detection	Classification	The Wikipedia article being edited.	32,439	EN
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	Question Answering	A natural language question.	200	EN, DE, FR, IT, PT, RO, ES, EU
WePS	Online reputation management	Document filtering	A set of tweets containing an organization name.	100	EN
WePS	Clustering Attribute	Document Clustering Information Extraction	A person name.	300	

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
	Extraction				
CriES	CriES Pilot Challenge	Expert Search	Questions posted to Yahoo! Answers	60	EN, DE, FR, ES
LogCLEF	LogCLEF	Log Analysis	Queries in logs can be seen as topics.	~4,000,000 of records	Any language

Table 13: Ground truth generation for the tasks in the CLEF 2010 Labs.

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
CLEF-IP	Patent Classification Prior Art Candidates Search	All documents in the collection	Automatic relevance assessments		
ImageCLEF	Medical image retrieval	Pooling (top 50 docs) ~30,000 topic/document pairs accessed	12	Medical doctors who are also students in biomedical informatics	~200 hours
ImageCLEF	Photo Annotation	All images in the collection	Unknown	Crowdsourcing via Mechanical Turk	The assessment was split in 4 surveys per image, answered by 3 turkers per image. Answering one survey took on average ~48 seconds.
ImageCLEF	Robot Vision	All images in the collection	Automatic relevance assessments		
ImageCLEF	Wikipedia image retrieval	Pooling (top 100 docs) ~186,000 topic/document pairs accessed	~10	Volunteers (organizers, participants, others)	~300 hours
PAN	Plagiarism Detection	All documents in the collection	Automatic relevance assessments		
PAN	Wikipedia Vandalism Detection	All documents in the collection	753	Crowdsourcing via Mechanical Turk	On average, a second per edit.

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
ResPubliQA	Paragraph Selection (PS) Answer Selection (AS)	All paragraphs and answers returned by participating system	Each run was first automatically compared against the manually produced Gold Standard. Any non-matching result was judged by native speakers.	The organizers	Variable
WEPS	Online reputation management	43,730 tweets	902	Crowdsourcing via Mechanical Turk	Variable
WePS	Clustering Attribute Extraction	60,000	~100	Crowdsourcing via Mechanical Turk	~30 seconds per document
CriES	CriES Pilot Challenge	Pooling (top 10 experts)	3	Students	150 hours
LogCLEF	LogCLEF			Some of the participants created a ground truth for themselves.	

Appendix V: Qualitative analysis of CLEF campaign website

This appendix discusses specific remarks for each of the seven characteristics and their sub-characteristics used for assessing the quality of the CLEF campaign website.

1. Architecture

- a. **Structure:** The structure of the website is adequate for maintaining information on the evaluation activities performed in CLEF. The informative structure mainly follows an evaluation campaign structure, specifically providing information on tracks in an edition, tasks in a track, and links to documents (paper and presentations) that describe the contributions of CLEF participants.

The navigation bar reported on the left of the homepage allows the navigation of the resources produced by the evaluation campaigns (e.g., publications) and content associated with the diverse CLEF editions; navigation is allowed only on a per edition basis, namely per year. Since the structure is strongly linked to an evaluation campaign workflow, it is easy to understand for a user who is familiar with information access and retrieval evaluation.

Different layouts are adopted in pages that display the same type of content (e.g. workshop program) across the different editions. An example is the working notes pages. The navigation bar on the left provides access to the different editions on a per year basis. For a subset of the editions, namely those from 2005 to 2009, the navigation bar on the left changes when accessing the working notes pages: the new bar provides access to papers in the working notes on a per track basis where the link corresponding to a track is an anchor to the part of the page where information on papers concerning the track is reported. This may cause an ambiguous interpretation of the left navigation bar. Links to different types of resources (e.g. PDF, Microsoft Word, Power-point or web pages) could be diversified.

- b. **Site map:** There is no map that describes the structure of the website. Since the structure is fairly simple, the absence of a map does not significantly affect the quality of the informative architecture.
- c. **Navigation:** Reaching track information and resources for a specific edition is simple because of the tree structure provided by the per year categorization of the contents. But the relationships between the resources, e.g. edition proceedings and the corresponding tracks and tasks, are not intuitive. In other words the current structure supports the user when adopting an “orienteering” access, but it does not support “teleportation” (e.g. from notebook or proceedings to the workshop or the tracks).

Actions:

- ⤴ Use of a uniform layout for pages that display the same type of information or the same type of resource.

- ⤴ Provide content access through categories other than the editions, i.e. not only on a per year basis.
- ⤴ Adopt a navigation bar to make pages of the main categories (i.e. editions, tracks, publications, resources) accessible from each page of the website.
- ⤴ Adopt a breadcrumb trail to help users understand the position of the current page in the context of the overall website structure.
- ⤴ Redesign the website structure in order to make explicit possible relationships between resources and content present in the website, e.g. by means of hyperlinks. This objective is pertinent to the PROMISE goal, i.e. helping users access information and knowledge that result from previous evaluation activities.

2. Communication

- a. **Home page:** The homepage currently reports a very brief description of the goal of CLEF. This description should be improved in order to make the CLEF objectives and achievements comprehensible for different types of user, e.g. also not expert in the field of information access and retrieval systems evaluation. The description as well as the structure of the website can be improved in order to guide the users across the content available.
- b. **Brand image:** The brand image can be adopted for allowing homepage access from any page in the website; indeed in some cases homepage access requires quite elaborate paths.
- c. **Graphics:** The target of the website is mainly experts in the field of information access and retrieval and therefore graphics has little importance. Despite this, the graphics should be improved because it can affect the use of the website, e.g. increasing the reading time because of the use of frames where external resources are displayed, or increasing the eyestrain because of the selected colors. Pages are usually too long; this was negatively perceived by the students when navigating the website. There are also several character formatting issues. Lastly, different choices in terms of fonts and page layout across the diverse years convey the idea that information and resources are provided from diverse sources.

Actions:

- ⤴ Report a more effective description on the homepage in order to help the user understand what kind of content they can find on the website.
- ⤴ Access to the homepage by a click on the brand image.

3. Functions

- a. **Adequacy:** The website does not provide specific functions but it mainly serves as an archive for information and resources of the diverse editions of CLEF. One of the functionalities requested by most of the students is a tool for full-text

search. Pages are static and the URL does not change when the user navigates across the pages, thus not allowing links to specific pages. A restricted area is present but there is no information on the procedure to access that area.

- b. **Correctness:** There is a number of character formatting issues. A number of links are broken and should be fixed or removed (63 broken links and 6 bad local links found by Xenu's Link Sleuth 1.3.8).

Actions:

- ✎ Add a full text search tool to facilitate information access and retrieval.
- ✎ Map each page to a specific URL that shows the position of the user within the website and that allows the page to be linked.
- ✎ Fix formatting issues and broken links.

4. Content

- a. **Categorization/labeling:** The categorization of content and resources is effective; it mainly follows the structure of an evaluation campaign and considers the type of resources that can result from an evaluation activity. This categorization is adopted to guide the user through the content of the website, specifically by means of an edition-based navigation; this approach can be extended, for instance, to also allow navigation on a per track basis. Even though resources and publications are categorized, this categorization is not used for navigation purposes. The list of entries in the "publications" section is quite long and a better management of the contents can be obtained by splitting this section into a set of subsections (e.g. journal papers, conference papers).
- b. **Style:** The style adopted is suitable for the purpose of the website and the type of users: the relevant information is provided. The homepage could provide a description of the objectives of CLEF and its website also for non-expert users, thus making clearer what kind of contents the website makes available.
- c. **Information:** The information reported on the website is perceived as reliable. The website update mainly involves pages for the new CLEF editions (workshop program, working notes, and link to proceedings). The "publications" page should be updated. Some links should be removed since they are no longer useful, e.g. the link to the submission page for the final version of the papers. "Publications", "Links", and "Archives" pages should report the date of the last update.
- d. **Localization:** All the content available in the website is in English. This choice is adequate since the type of users of this website is more likely to be researchers or experts in the field of information access and retrieval.

The animated gif image at the bottom on the navigation bar seems to suggest that the content is available in diverse languages and therefore it can be misleading.

Actions:

- ⤴ The animated gif should be substituted with a different visualization that is not misleading, i.e. that does not suggest that the website is available in different languages.
- ⤴ Information that is no longer useful should be removed.
- ⤴ Categorization should be adopted to support navigation not only on a per edition basis; publication categorization should also be exploited to make content access easier.
- ⤴ The homepage should report a description, which is comprehensible for non-expert users.

5. Management

- a. **Availability:** The website was always available during the period when the students performed their analysis.
- b. **Monitoring:** Two types of monitoring activities were considered for the analysis. The first type of activity is that performed by the webmaster in order to check possible issues in the website, in term of webpage visualization and broken links. Some links are still broken, thus suggesting that this type of monitoring activity is not regularly performed. The second type of activity concerns the number of accesses, registration and login to the website. This information is not maintained or at least not made publicly available.
- c. **Update:** The website is updated several times a year for a new edition of the CLEF campaign. News is currently reported on the homepage; a dedicated page or section for the news can improve the user perception that the website is updated. The date of the last update is not visualized on the website pages.
- d. **Relationship with the users:** A contact page is available to gain more information on the campaign or to be included in the CLEF mailing list. There is no forum or tool to make the interaction among participants easier.

Actions:

- ⤴ A monitoring tool should be added to provide information on the number of accesses to the website and the diverse resources made available through it; this could also be useful for measuring the impact of the evaluation activity on multimodal and multimedia information access.
- ⤴ The content update and the monitoring activity could be performed more frequently.

6. Accessibility

The website design did not take into account its use by impaired people.

7. Usability

- a. **Effectiveness:** The user looking for information on the evaluation activities carried out in CLEF can completely reach his objective.
- b. **Efficiency:** The navigation is strictly related to the evaluation activity workflow, thus making navigation more difficult for users unfamiliar with the campaign. The possibility to access resources also by means of other categories can reduce both user effort and the time required to reach a specific content; efficiency could be also improved by providing users with additional information on the structure and the content that can be found in the website. Another source that may negatively affect user perception of the website is the visualization of external content within website frames. This is the case of the proceedings pages. This was negatively perceived by most of the students, since it reduces space and reading capability.
- c. **User satisfaction:** The user satisfaction in general is quite low. The main causes are (i) the difficulty in content navigation, mainly based on a per year categorization, (ii) the fact that users have difficulty in understanding the exact position within the website and accessing the homepage from each page, and (iii) the lack of visual difference among the links to the diverse types of resources.

Actions:

- ⤴ The navigation should be made easier.
- ⤴ The connections between diverse informative resources and their differentiation should be made explicit.

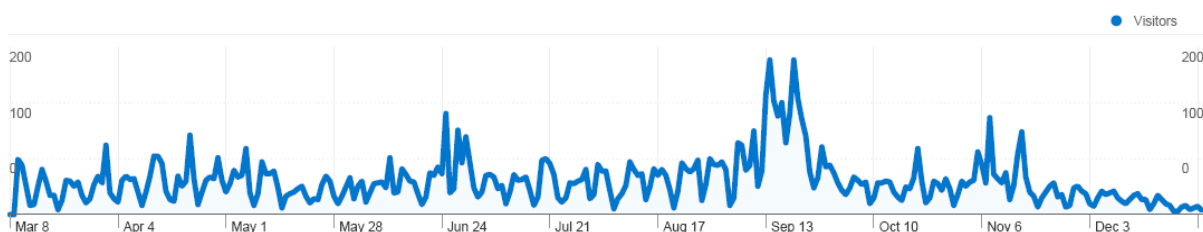
Appendix VI: CLEF 2010 website statistics

The high visibility of the CLEF websites is evident in the number of visits they attract. The CLEF 2010 website had 14,498 visits and 6,861 visitors in 2010, coming especially from Italy, whereas it still has a good number of visitors in 2011, and quite surprisingly, they mostly come from the United States.

1. CLEF 2010 website: March 2010 – December 2010 statistics

Visitors Overview

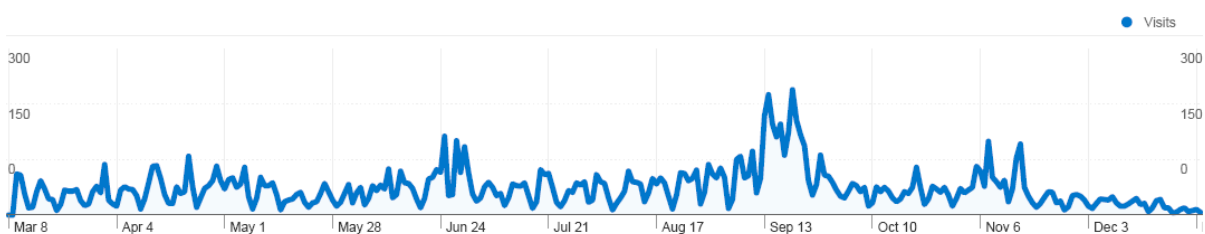
Mar 8, 2010 - Dec 31, 2010
Comparing to: Site



www.clef2010.org

Traffic Sources Overview

Mar 8, 2010 - Dec 31, 2010
Comparing to: Site



All traffic sources sent a total of 14,498 visits

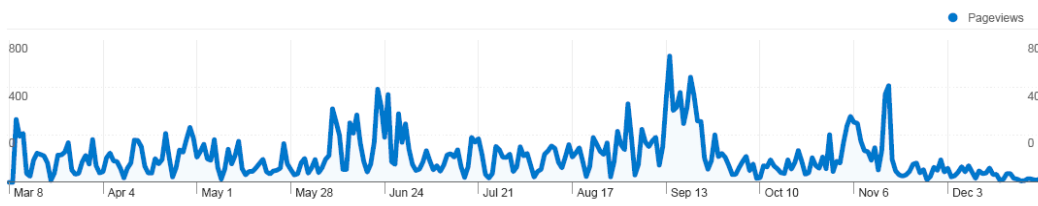
24.56% Direct Traffic
30.94% Referring Sites
44.50% Search Engines



Search Engines
6,452.00 (44.50%)
Referring Sites
4,485.00 (30.94%)
Direct Traffic
3,561.00 (24.56%)

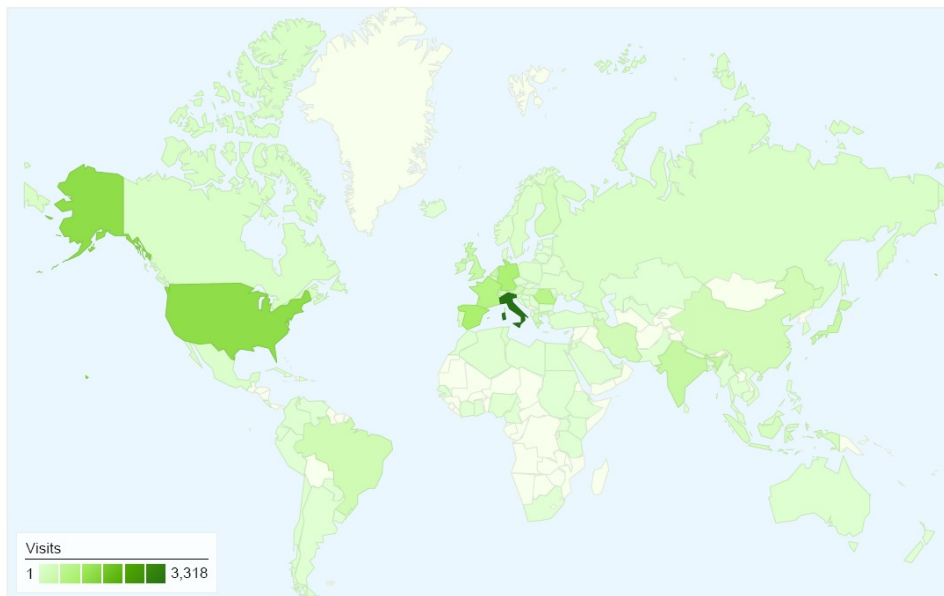
Content Overview

Mar 8, 2018 Dec 31, 2018
Comparing to: Site



Map Overlay

Mar 8, 2018 Dec 31, 2018
Comparing to: Site

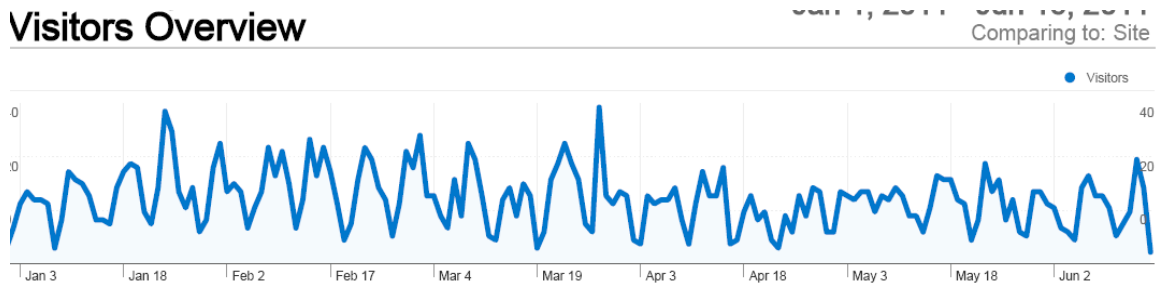


14,498 visits came from 111 countries/territories

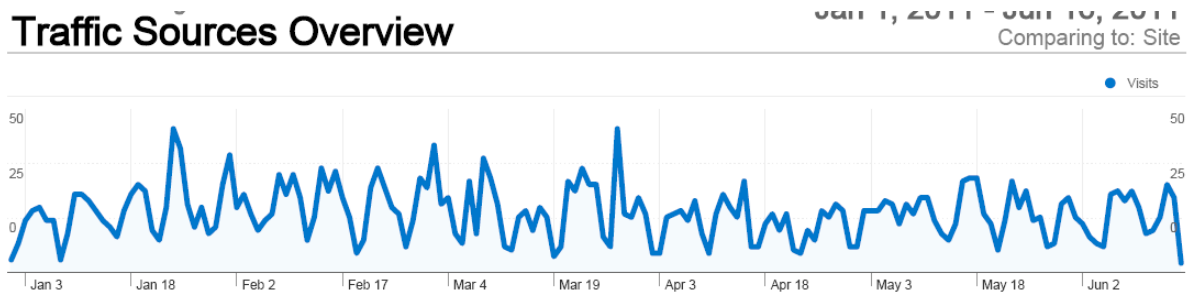
Site Usage					
Visits 14,498 % of Site Total: 100.00%	Pages/Visit 2.98 Site Avg: 2.98 (0.00%)	Avg. Time on Site 00:02:49 Site Avg: 00:02:49 (0.00%)	% New Visits 46.97% Site Avg: 46.87% (0.21%)	Bounce Rate 45.06% Site Avg: 45.06% (0.00%)	
Country/Territory	Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
Italy	3,318	3.98	00:04:06	38.37%	44.15%
United States	1,398	2.37	00:01:59	60.66%	55.36%
Spain	1,013	3.14	00:02:48	37.81%	33.96%
Germany	939	2.78	00:02:26	45.26%	42.17%
France	799	3.08	00:03:15	47.81%	38.42%
Romania	737	2.45	00:01:54	41.66%	51.70%
India	490	2.76	00:02:50	61.63%	38.98%
United Kingdom	458	2.44	00:01:41	62.01%	51.31%
Japan	418	2.90	00:02:43	40.43%	47.13%

2. CLEF 2010 website: January 2011 – June 2011 statistics

Visitors Overview



Traffic Sources Overview



All traffic sources sent a total of 3,089 visits

18.45% Direct Traffic

43.48% Referring Sites

38.07% Search Engines



■ Referring Sites
1,343.00 (43.48%)

■ Search Engines
1,176.00 (38.07%)

■ Direct Traffic
570.00 (18.45%)

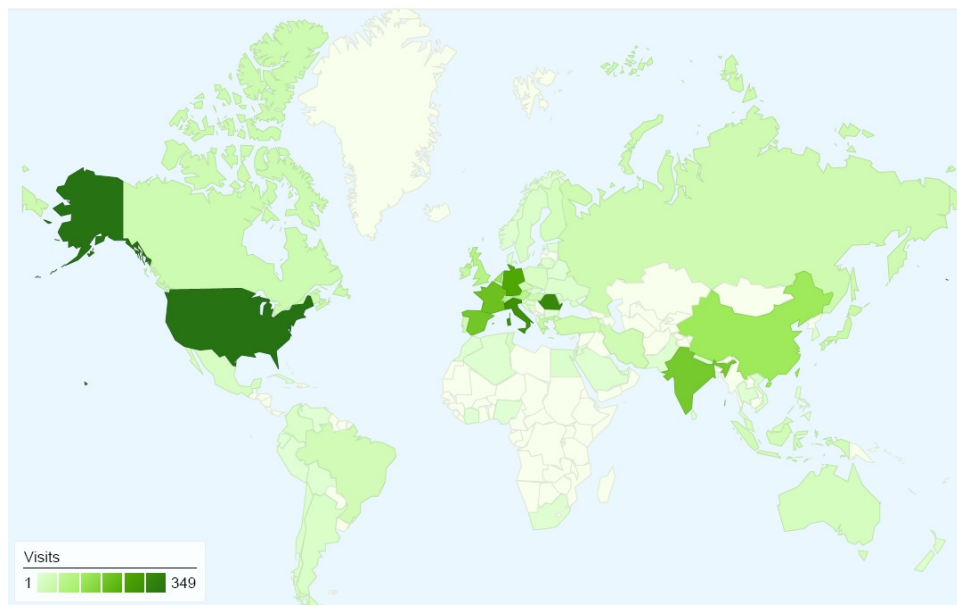
Top Content

Pages	Pageviews	% Pageviews
/	1,963	29.09%
/index.php?page=pages/proceedings.php	1,461	21.65%
/index.php?page=pages/labs.html	504	7.47%
/index.php?page=pages/schedule.html	447	6.62%
/index.php?page=pages/acceptedPapers.html	441	6.53%

Map Overlay

Jan 1, 2011 - Jan 10, 2011

Comparing to: Site



3,089 visits came from 80 countries/territories

Site Usage

Visits 3,089 % of Site Total: 100.00%		Pages/Visit 2.18 Site Avg: 2.18 (0.00%)		Avg. Time on Site 00:01:40 Site Avg: 00:01:40 (0.00%)		% New Visits 57.62% Site Avg: 57.40% (0.39%)		Bounce Rate 57.95% Site Avg: 57.95% (0.00%)	
Country/Territory			Visits	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate		
United States			349	2.10	00:01:28	65.33%	58.45%		
Romania			305	1.59	00:01:30	31.15%	79.02%		
Italy			286	1.97	00:01:13	50.00%	56.64%		
Germany			238	2.66	00:02:13	59.24%	48.32%		
France			195	2.42	00:01:44	50.77%	54.87%		
Spain			186	2.42	00:01:33	50.54%	48.92%		
India			181	1.88	00:01:46	61.33%	55.80%		
Netherlands			130	2.41	00:02:00	43.85%	40.77%		
China			127	2.71	00:03:12	65.35%	61.42%		