PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 6.2
Report on the outcomes of the second year evaluation activities

Version 1.0, 26 June 2012

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

PROMISE
Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

SEVENTH FRAMEWORK PROGRAMME

Document Information

Deliverable number:	D6.2
Deliverable title:	Report on the outcomes of the second year evaluation activities
Delivery date:	31/08/2012
Lead contractor for this deliverable	HES-SO
Author(s):	Florina Piori, Vivien Petras, Maria Gäde, Birger Larsen, Theodora Tsikrika, Alba G. Seco de Herrera, Henning Müller
Participant(s):	All
Workpackage:	WP6
Workpackage title:	Evaluation activities
Workpackage leader:	HES-SO
Dissemination Level:	PU – Public
Version:	1.0
Keywords:	Evaluation activities, CLEF conference, CLEF Labs, CLEF-IP, ImageCLEF, CHiC

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	05/12/11	Draft	Alba G. Seco de Herrera and Theodora Tsikrika (HES-SO)	First draft
0.2	31/05/12	Draft	All authors (several partners)	Inclusion of work of the various partners
0.3	01.06.2012	Draft	Henning Müller (HES-SO)	Reworked introduction and text
0.4	05.06.2012	Draft	Alba G. Seco de Herrera and Henning Müller (HES-SO)	Sent for internal review
1.0	25./26.06.2012	Final	Alba G. Seco de Herrera and Henning Müller (HES-SO)	Revised after internal review

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191

Abstract

This deliverable reports on the outcomes of the evaluation activities (in WP6) in the second 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 and workshop. 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., unlocking culture, search for innovation and visual clinical decision support. 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. We compare this year with the first PROMISE year outcomes. The deliverable concludes with an outlook on the evaluation activities for the third year of PROMISE.

Table of Contents

Executive Summary

This deliverable presents the main outcomes of the evaluation activities in the second year of PROMISE, i.e., the outcomes of 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 case domains. The deliverable concludes with an outlook on the evaluation activities for the third year.

- **Evaluation activities in CLEF 2011: 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. Since 2010, CLEF has consisted 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 2011 conference** together with a short description of the **CLEF 2011 labs** and the participation to them. To gain insights on the outcomes of the CLEF 2011 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 the questionnaires sent to the CLEF 2011 lab organizers. These results can be summarized as follows:

1. **Tasks**: A total of 18 tasks were investigated in the CLEF 2011 labs, four more than in CLEF 2010: eight classification tasks and six (ad-hoc) information retrieval tasks and the combination of both in the medical task. There were also two question answering tasks and one log analysis task.
2. **Main advancements:** The main difference between the CLEF 2010 and CLEF 2011 labs is the considerable number of new tasks that were introduced. Similar to the year before, 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.
3. **Main trends in the participants' approaches**: Given the high heterogeneity of the tasks, the main purpose of the analysis in this deliverable was to monitor the trends over each task. To this end, we compare the main trends and experimental outcomes between the CLEF 2011 and the CLEF 2010 labs for the tasks that ran during both evaluation campaigns.
4. **Main problems**: The main problems reported by lab organizers concern (i) availability and quality of underlying infrastructures to support their evaluation activities, such as annotation systems, experiment submission systems, and collaborative systems for enabling efficient communication among participants and organizers, (ii) low participation rate compared to the number of registrations, particularly for the CLEF-IP classification tasks and MusiCLEF lab; (iii) difficulties in creating a realistic test collection and in providing additional resources to support participants in their experimentation. PROMISE aims to address these challenges through the development of the PROMISE evaluation infrastructure and by promoting evaluation tasks that correspond to well-defined and compelling use cases.
5. **Test collections generated by the CLEF 2010 labs**

a. **Collections:** The CLEF 2011 Labs employed a total of 18 collections for the 17 tasks. The overall trend is towards a continuous growth of the labs, both with new tasks and new collections being introduced. The continuous update of existing datasets manifests a tendency to increase the volume of data and include multilingual aspects. The collections described in this deliverable are evidence of the large size of the datasets employed in the PROMISE evaluation activities.
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 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 in this deliverable.
c. **Ground truth:** Compared to 2010, the same number of tasks used crowdsourcing for creating the human relevance assessments, but more tasks relied either on automatically generated relevance assessments or on human assessors. In this latter case, the human effort required to generate relevance assessments varies greatly based on the nature and difficulty of the task, but can reach up to several weeks for a single task. 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.

- **Evaluation activities for PROMISE use cases**

We then present the outcomes of 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.

1. **"Visual Clinical Decision Support" Use Case** (Medical retrieval task at ImageCLEF lab)
    a. *ImageCLEFmed was the most popular lab*. The number of registrations reached a new maximum with 60, 9 more than in CLEF 2010.
    b. *More research is necessary for the effective and robust combination of evidence from various modalities.* Similar to 2010, combination of evidence from various modalities is the most effective approach for the modality detection and medical image retrieval tasks, whereas textual methods are most effective for the medical case retrieval task. Extending the training data lead to best results in modality classification.
2. **"Search for Innovation" Use Case** (CLEF-IP Lab)
    a. *There is a need to reformulate The Prior Art Candidates Search.* Taking a closer look at the results and methods obtained in this task there is a need to at least reformulate this task such as to focus on specific tasks of a patent professional. Neither of these is currently reflected in the CLEF-IP Lab.

b. *Very good scores obtained in the classification tasks*. As in 2010, seem to show that patent classification, at least up to the subclass level of the IPC system, is an easy task. A similar task will not be organized in 2012.
c. *More research is necessary for the effective and robust combination of evidence from various modalities*. It appears that the big majority of research groups, in order to obtain good results, work either on text processing, or on image processing but not on both. We have also recognized that the set up of the image-based patent retrieval task made it difficult to tackle. Breaking this task in finer-grained (sub) tasks involving image processing may be a better way to approach patent retrieval using patent images.
d. *Involvement of patent professionals with this use case contributed to motivating CLEF-IP Lab and PatOlympics participants*. It provided feedback to the work done in the use case. Concurrently, members of the EPO have recognized the relevancy of the research within the CLEF-IP Lab to more readily provide expanded support in 2012.

3. **"Unlocking Culture" Use Case** (CHIC Lab)
   a. *CHiC as a workshop*. In 2011, the CHiC2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage workshop investigated evaluation efforts in the cultural heritage field as well as defining user scenarios and identifying possible relevant metrics for a benchmark CLEF lab.
   b. *CHiC as a lab.* CHiC 2012 pilot evaluation lab aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems, creating evaluation tasks that represent the particular contingencies of the cultural heritage domain and should support system developers in defining systematic evaluation standards.

1 Introduction

PROMISE is working on providing a virtual laboratory for conducting participative research and experimentation to carry out, advance and bring automation into the evaluation and benchmarking of such complex information systems, by facilitating management and offering access, curation, preservation, re-use, analysis, visualization, and mining of the collected experimental data. To this end, PROMISE organizes CLEF, an experimental evaluation activity for multilingual and multimedia information systems at an international level and on an annual basis. CLEF consists of an independent peer-reviewed conference on a broad range of topics in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test various aspects of mono and cross-language Information retrieval systems.

This deliverable reports on the outcomes of the concrete experimental evaluation activities that have taken place during the second year of PROMISE, with particular focus on the evaluation campaigns organized for the three domains of the PROMISE use cases, i.e., unlocking culture, search for innovation and visual clinical decision support.

Comparisons between CLEF 2010 that was organized in year 1 of PROMISE and CLEF 2011 are performed based on the material in PROMISE Deliverable 6.1 [1]. The state of the activities and labs for CLEF 2012 will also be mentioned but results will only be available in the third PROMISE evaluation report.

This deliverable is structured as follows. Section 2 provides an overview of the second year evaluation activities by discussing the main outcomes of and the lessons learned from the CLEF 2011 conference and labs. Section 3 focuses on one of the main outcomes of these experimental evaluation activities, the CLEF 2011 lab test collections. Sections 4, 5, and 6 provide a more detailed analysis of the outcomes of the evaluation activities for the three PROMISE Use Cases. Section 7 presents the new task impact analysis for the CLEF initiative. Section 8 concludes by providing an outlook on the current status of the CLEF 2012 conference and labs.

Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191

2 Overview of the second year evaluation activities

2.1 CLEF 2011 Conference and Labs

Following the success of the new model of the CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation, CLEF 2011, as an activity of PROMISE, was organised in a similar way. CLEF 2011 consisted of an independent conference on a broad range of questions in the fields of multilingual and multimodal information access evaluation and a set of labs that continued the CLEF tradition of community-based evaluation.

The CLEF 2011 conference on Multilingual and Multimodal Information Access Evaluation was held in Amsterdam from 19th to 22nd September, 2011. For further information about CLEF 2011 conference, see [2].

2.1.1 CLEF 2011 Conference

The CLEF 2011 labs continued the CLEF tradition of community-based benchmarking and complemented it with workshops on emerging topics in evaluation methodology. Following the format introduced in 2010, two forms of labs were offered: labs could either be run as benchmarking activities "campaign-style" during the ten month period preceding the conference, or as "workshop-style" labs that could explore possible benchmarking activities and provide a means to discuss information retrieval evaluation challenges from various perspectives. There were 9 lab proposals: 6 were accepted as benchmarking (campaign-style) labs and 1 was accepted as a workshop, resulting in an acceptance rate of 7/9 (=77%), similar to 2010.

Table 1 lists the CLEF 2011 labs and the tasks organised within each of them. Compared to 2010:

- four benchmarking labs (CLEF-IP, ImageCLEF, PAN, QA4MRE[1]) returned;
- a workshop-style lab (LogCLEF) became a benchmarking lab;
- a new benchmarking lab (MusiCLEF) was introduced;
- a new workshop-style lab (CHiC) was introduced;
- two CLEF 2010 labs did not return: the WePS (Web People Search) benchmarking lab and the CriES (multi-lingual expert search in social media environments) workshop-style lab.

[1] QA4MRE is a continuation of the ResPubliQA CLEF 2010 benchmarking lab and other past CLEF tracks on question answering.

Table 1: CLEF 2011 benchmarking and workshop-style labs. New labs and tasks compared to 2010 are marked with a (*). Benchmarking labs that were previously workshop-style labs are marked with a (†).

Benchmarking labs, their tasks, and subtasks		
CLEF-IP	Patent Classification	
	Patent Image-based Classification*	
	Patent Image-based Prior Art Search*	
	Prior Art Candidates Search	
	Refined Patent Classification*	
ImageCLEF	Medical Image Classification and Retrieval	
	Photo Annotation	Photo Annotation
		Concept-based Photo Retrieval*
	Plant Identification*	
	Wikipedia Image Retrieval	
LogCLEF†	Multilingual Log File Analysis	
MusiCLEF*	Music Categorisation*	
	Music Identification*	
PAN	Authorship Identification*	
	Plagiarism Detection	
	Wikipedia Vandalism Detection	
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation*	
	Question Answering for Machine Reading Evaluation*	
Workshop-style Lab		
CHiC*	Cultural Heritage*	

Here is a brief description of the CLEF 2011 benchmarking labs:

1 **CLEF- IP:** a benchmarking activity on retrieval in the intellectual property domain [1], running since 2009. There were five tasks in 2011: the *Prior Art Candidates Search* task for finding prior art patent documents, i.e., finding patent documents that may invalidate a given patent application, the *Patent Classification* task for classifying documents into the subclass level of the International Patent Classification (IPC) hierarchical scheme, the *Refined Patent Classification* task, a refined version of the previous task that required classification into IPC levels deeper in the hierarchy than the subclass level, the *Patent Image-based Classification* task for classifying patent images into pre-defined image-related categories, and the *Patent Image-based Prior Art Search* pilot task for finding prior art patent documents given a patent application based on both textual and visual content.

2 **ImageCLEF:** a benchmarking activity on the cross–language annotation and retrieval of images, running since 2003. Four tasks were offered in 2011: the *Medical Image Classification and Retrieval* task [2] that used a data collection from the scientific literature for the classification of images according to their acquisition modality and

the retrieval of images or relevant cases given a medical professional's multimedia and multilingual information need, the *Photo Annotation* task [3] for the automated annotation of Flickr images with visual concepts and for the concept-based retrieval of such images, the *Plant Identification task* [4] for plant species identification based on leaf images, and the *Wikipedia Image Re*trieval task [5] for the multimodal and multilingual information retrieval from a collection of Wikipedia images.

3 **LogCLEF:** a benchmarking activity on *Multilingual Log File Analysis* [6], running since 2009. It focuses on the analysis of transaction logs and questions of language identification, query classification, and query drift, with the ultimate aim to gain insights into users' search behavior in multilingual contexts.

4 **MusiCLEF**: a benchmarking activity on music access and retrieval from real public music collections [7], introduced in 2011. Its major focus was on professional users and on the multimodal classification and retrieval of music through the combination of content-based and context-based evidence. Two tasks were offered in 2011: *Music categorisation* (auto-tagging) for categorizing music based on its possible usage in various scenarios and *Music Identification* for the identification, given a song, of versions (covers) of the same song.

5 **PAN:** a benchmarking activity on Uncovering Plagiarism, Authorship, and Social Software Misuse, running at CLEF since 2010. Three tasks were offered in 2011: the *Plagiarism Detection* task [8] for automatically detecting plagiarism, i.e., the act of copying another author's text and claiming its authorship, the *Author Identification* task [9] for determining the authorship of anonymous documents based on internal evidence, and the *Wikipedia Vandalism Detection* task [10] for automatically detecting when a Wikipedia article has been changed with malicious intent.

6 **QA4MRE:** a benchmarking activity on question answering for machine reading evaluation, a major innovation of past Question Answering (QA) tracks at CLEF. Two tasks were offered in 2011: the *Question Answering for Machine Reading Evaluation* task [11] for the identification of the answers to a set of multiple-choice questions given a single document as input, and the *Annotating Modality and Negation for a Machine Reading Evaluation* pilot task [12] for determining the machine's ability to understand extra-propositional aspects of meaning like modality and negation.

In summary, the CLEF 2011 benchmarking labs consist of a total of 17 tasks listed in Table 1: 7 of them are Classification tasks, 5 of them are (ad-hoc) Information Retrieval tasks, 2 of them contained both classification and information retrieval subtasks[2], 2 of them are Question Answering tasks, and 1 is Log Analysis. Compared to 2010, there is an increase both in the number of benchmarking labs (6 vs. 5) and in the number of their tasks (17 vs. 11), but these tasks are now more focused. All types of tasks offered in 2011, i.e., classification, information retrieval, question answering and log analysis, were also offered in 2010. Other types of tasks such as document filtering, document clustering, information extraction and expert search were only covered explicitly in 2010.

The following workshop-style lab was also held at CLEF 2011:

[2] These are the ImageCLEF medical image classification and retrieval and the ImageCLEF photo annotation tasks.

- **CHiC:** this workshop aimed at surveying use cases for information access to cultural heritage materials and review evaluation initiatives and approaches in order to identify future opportunities for novel evaluation experiments and measures [13]. Workshop participants were asked to introduce their ideas for possible evaluation scenarios resulting finally in a benchmarking lab for 2012.

The results of the experiments conducted within the CLEF 2011 labs were presented and discussed as sessions of half a day, one full day or one and a half days at CLEF 2011, on 19-22 September, in Amsterdam, The Netherlands. These sessions were run in parallel covering three of the four days of the CLEF Conference. A general poster session was arranged at the end of the first day, where all participants from all the 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.

2.1.2 Participation in the CLEF 2011 Labs

The CLEF 2011 evaluation activities have achieved high visibility. Out of the 169 research groups initially registered, a total of 95 institutions from 30 countries participated in the benchmarking activities by submitting a total of 787 runs, with 89 participants also preparing reports of their experimental results that were published in the CLEF 2011 Labs and Workshop Notebook papers [1]. Compared to 2010, there is a slight decrease both in overall number of registrations (216 VS. 169), participations (95 vs. 110), and per lab participations (CLEF-IP 8 vs. 19, ImageCLEF 44 vs. 49, PAN 24 vs. 31, QA4MRE 13 vs. 24, LogCLEF 16 vs. 19, MusiCLEF 2). There is a significant increase though in the overall number of submissions (787 vs. 595), although the average number of submissions per task remained the same at 49[3]. Table 12 in

Appendix II: Participation in the CLEF 2011 labs provides a more detailed breakdown on the number of registrations, participations, return participations per task, and submitted runs per task.

Similarly to 2010, the most popular CLEF 2011 Lab was ImageCLEF, which was able to attract most registrations and participations, not only from Europe but also from the United States and other countries. The most established tasks, i.e., those running for a number of years such as the ImageCLEF medical and photo annotation tasks, attracted the most registrations and participations. Another well-established task, the Question Answering for Machine Reading Evaluation task at QA4MRE that continues a long history of question answering tasks at CLEF, also attracted a considerable number of registrations and participations, equal to those of 2010. Among the newcomers, the Authorship Identification task at PAN and the Plant Identification task at ImageCLEF attracted significant numbers both of registrations and participations, whereas the new tasks at CLEF-IP and the newly introduced MusiCLEF lab did not manage to achieve that. A possible explanation for the latter is that the interdisciplinary nature of the new CLEF-IP tasks rendered them quite difficult to tackle and probably required expertise outside of the community already

[3] The averages are computed by considering that there were 12 tasks that accepted submissions in 2010 and 16 in 2011.

established at the lab. Similarly for MusiCLEF, it appears that other evaluation forums might be more suitable for attracting participants with the required expertise. On the other hand, more established labs, such as ImageCLEF and PAN, manage to attract participants for their new tasks probably due to their existing infrastructure and knowledge on supporting new tasks and also the wide-reaching research communities they have built over the years.

The participation rate (i.e., the number of registered research groups that actually submitted their results to the lab) is on average 25% with the highest for the Prior Art Candidates Search task at CLEF-IP (9 participants out of 17 registrations), and the lowest for the Music Categorisation task at MusiCLEF that did not attract any participants despite having 20 registrations, and the Annotating Modality and Negation for a Machine Reading Evaluation pilot task QA4MRE that had no registrations or participations. These two outliers are also partly responsible for the decrease in the participation rate compared to 2010.

Return participations from the previous year are on average around 50%, an increase compared to 2010 when it was 40%, indicating that a large number of researchers rely year after year on the resources created in the context of the CLEF evaluation activities. In particular, all the tasks that ran in previous CLEF editions had return participants, apart from the Wikipedia Vandalism Detection task at PAN that had no return participants. It is worth noting that for some tasks, the number of return participants is extremely high, such as the Question Answering for Machine Reading Evaluation task at QA4MRE with 100% return participation and the Wikipedia image retrieval task at ImageCLEF with 82%.

The number of submissions varies greatly per task, with an average of 49. Notable cases are the ImageCLEF Medical task with 207 submissions, an increase of 33% compared to 2010, the plagiarism detection task at PAN with 105 submissions, an almost 300% increase compared to 2010, and the authorship identification task at PAN that attracted 92 submissions in its first year.

Furthermore, each lab employs a submission system, similarly to 2010, indicating that the need for the provision of a unified evaluation environment and infrastructure, as the one currently developed in PROMISE. Nevertheless, there are steps towards this direction, with the CLEF-IP lab using the PROMISE evaluation infrastructure in 2011, whereas it is foreseen that the ImageCLEF lab, and in particular its medical task, will use it in 2012.

2.2 Main advancements

The main difference between the CLEF 2010 and CLEF 2011 labs is the considerable number of new tasks that were introduced. Only half the tasks remained the same, while QA4MRE presented a major innovation of the long-running Question-Answering CLEF lab. Table 13 in Appendix III: Main outcomes of the CLEF 2011 Labs, presents the main differences between the two years as pointed out by the task organizers.

Many of the tasks (4 out of the 8 tasks that also ran in 2010 and the QA4MRE tasks) employed larger collections, either by updating existing collections or creating new ones from scratch. Particular efforts were made towards making the tasks more realistic by improving not only the collections, e.g., the efforts made by PAN Plagiarism Detection to make the collection more realistic and therefore more difficult), but also the topic development process, e.g., by increasing the number of topics, by creating topics that correspond more closely to real practice (CLEF-IP, ImageCLEF, and PAN) and by improving

the language distribution.

Similar than in 2010, 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 and experimental outcomes

Table 14 in Appendix III: Main outcomes of the CLEF 2011 Labs presents the main trends among the participants' approaches, 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 monitor the trends over each task. To this end, we compare the main trends and experimental outcomes between the CLEF 2011 and the CLEF 2010 labs for the tasks that ran during both evaluation campaigns.

1. *CLEF-IP Patent Classification*: The two participants in the tasks have used various solutions to the classification problem, both with very good results. A linguistic classification system was used to implement three classifiers. Combination of retrieval and classification algorithms was also applied. The first participant chose to treat only English language topics were used as well as an external service to translate all non-English topics into English.
2. *CLEF-IP Prior Art Candidates Search*: Most of the participants focus on applied linguistic methods to process the data and analysing how these impact the search results. Some participants treated only the English set of topics but also cross-lingual search was applied translating queries into English. Metadata was used to construct better queries, restrict the search space, or filter out retrieval results.
3. *ImageCLEF Medical Image Classification and Retrieval*: The main trend in both CLEF 2010 and 2011 was mapping of text onto medical ontologies such as MeSH[4] (Medical Subject Headings) terms. In 2011, the MeSH hierarchy was also used and query expansion was often successful. Similarly to CLEF 2010, in CLEF 2011 visual approaches have good early precision. Even if fusion is hard to do multimodal approaches were often best.
4. *ImageCLEF Photo Annotation*: Both in CLEF 2010 and in CLEF 2011, most participants relied on Scale-Invariant feature transform (SIFT) [3] and discriminative approaches and also the results in using multimodal evidence outperform classification with single modality information. Given that several textual approaches were applied in CLEF 2011, different from CLEF 2010, their performance could be better analyzed. The results indicated that the performance of textual runs was close to that of the visual runs.
5. *ImageCLEF Wikipedia Image Retrieval*: The main CLEF 2010 trends of having more multimodal (and multilingual approaches) being applied and having many groups use external sources to enhance retrieval (e.g., Flickr, WordNet etc.) continued also in CLEF 2011. In addition, components provided by participating groups were re-used by several participants so that each group could focus only on the specific

[4] http://www.nlm.nih.gov/mesh/

aspects of research that were of interest to them. Similarly to CLEF 2010, in CLEF 2011 multilingual approaches were more successful than monolingual ones and for the majority of the participants that submitted both multimodal and monomedia approaches, their multimodal outperformed mono-media runs. This is probably due to the increased number of visual examples, improved visual features, and more appropriate fusion techniques.

6. *LogCLEF*: In CLEF 2011, the task consolidated further through the generation of ground truth and sharing of resources, and thus addressing some of the issues raised in CLEF 2010. As a result, measurable effects on the success of search query based on language correlations could be observed, with native language vs. interface language influencing how users interact with the application. Also, interface language changes during a session may give hints about user search preferences.
7. *PAN Plagiarism Detection*: The exhaustive comparison of suspicious documents to source documents approach applied in CLEF 2010 was also applied in CLEF 2011, together with other techniques (e.g., similar document indexing pipeline, dotplot-based plagiarism extraction, and intrinsic detection based on outlier detection). This led to remarkable performance improvements for intrinsic detection, whereas external detection posed a renewed challenge with few participants managing to perform well on all measures due to the more difficult corpus being used.
8. *PAN Wikipedia Vandalism Detection*: In CLEF 2010, various paradigms for features have been employed, some content- based, some context-based, but no participant employed two paradigms at the same time. This changed in CLEF 2011, where the various paradigms were combined and the best performance was achieved with a combination of content-based and context-based features. Also, in CLEF 2011, it was the first time that language-dependent features and a-posteriori features were applied. As a result, detection performance improved significantly as a result of the new kinds of features employed.

2.4 Main problems from an organizational point of view

The main problems reported by lab organizers concern 1) the availability and quality of the underlying infrastructure to support their evaluation activities, such as annotation systems, experiment submission systems, and collaborative systems for enabling efficient communication among participants and organizers, 2) the low participation rate compared to the number of registrations, particularly for the CLEF-IP classification tasks and MusiCLEF lab, and 3) the difficulties in creating a realistic test collection and in providing additional resources to support participants in their experimentation.

PROMISE addresses the issue of the availability of appropriate infrastructures through the development of the PROMISE evaluation infrastructure, already used by the CLEF-IP lab in 2011 and with further labs adopting it in 2012. Furthermore, PROMISE can also contribute towards the increase of the participation rate by promoting evaluation tasks that correspond to well-defined and compelling use cases, and thus stimulate research and development in the related fields. The framework for developing these use cases and evaluation tasks can

also guide lab organizers in building more realistic test collections.

3 Outcomes of evaluation activities: CLEF 2011 lab test collections

3.1 Collections

The CLEF 2011 Labs employed a total of 18 collections for the 17 tasks; 3 tasks in the CLEF-IP lab shared the same collection, whereas the ImageCLEF photo annotation task employed 2 separate collections, one for each of its subtasks, and LogCLEF employed 3 collections. A description of each collection and some statistics are presented in Appendix IV: CLEF 2011 Labs Test Collections.

The employed collections have either been purpose-built for the labs or have been extracted from already existing collections. Twelve of them (66%) were completely new and employed for the first time in 2011, an increase compared to 2010 when 53% of the collections were completely new. The remaining six collections have been used once or at most twice before in previous years of the same labs. Out of these previously used collections, half have remained unchanged over these couple of years (MIR Flickr collection for the ImageCLEF Photo Annotation: Annotation subtask, ImageCLEF Wikipedia collection, and LogCLEF DBS logs) and half have been updated mainly through the addition of new documents (CLEF-IP 2011, CLEF-IP-IMG 2011, and LogCLEF TEL logs). The overall trend is towards a continuous growth of the labs, both with new tasks and new collections being introduced.

Thirteen of the collections (72%) are multilingual, ranging from two to five languages, with the exception of the LogCLEF collections that can in principle cover any language. The monolingual collections include three of the ImageCLEF collections, given that they focus on multimedia retrieval and its language independent nature, and the PAN Authorship Identification and QA4MRE-modality corpora employed in newly-introduced tasks that aimed to keep the level of complexity down during their first year. Compared to 2010 when about half of the collections were multilingual, there is a clear trend towards increasing the multilinguality of the employed collections.

The size of the collections and the number of documents they contain vary widely, with the size ranging between 148 KB and 1.5TB and the number of documents between 12 and 3.5 millions. The overall trend appears to be towards larger collections with a size of several gigabytes being the norm. All tasks that used new collections in 2011 employed larger collections compared to 2010 with notable examples the QA4MRE collection that was 86 times larger than the collection used in 2010 and the ImageCLEF medical task collection that was three times larger.

The collections described in this section are evidence of the large size of the datasets employed in the PROMISE evaluation activities. The continuous update of existing datasets manifests a tendency to increase the volume of data and include multilingual aspects.

3.2 Topics

The nature and the number of topics employed in the tasks of the CLEF 2011 labs depend on the type of the task and are described in Table 16 in Appendix IV: CLEF 2011 Labs Test

Collections.

In the classification tasks, the documents to be classified range from a few hundred (e.g., 100, 380, or 400) to a few thousand (e.g., 1,000 or 3,000) up to several thousand (e.g., 10,000 or 30,000). In most cases, the classes are less than 100, with the exception of the CLEF-IP patent application classification tasks that consider several hundreds or even several thousands of classes. A difference compared to 2010 when the largest number of classes considered was less than 1,000, there is a new task (CLEF-IP Refined Patent Classification task) that evaluated classification into several thousands of classes. Overall, though, the characteristics of the classification tasks have largely remained the same, with the number of classes being determined to a large extent by the effort required for generating the ground truth. The retrieval tasks range between 30-50 topics for ImageCLEF tasks to 619 for the Music Identification task up to a few thousand topics for the CLEF-IP prior art search and the plagiarism detection tasks. This is similar to CLEF 2010. It is worth noting that the overwhelming 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

Table 17 in Appendix III: Main outcomes of the CLEF 2011 Labs briefly presents the process for the ground truth generation followed in each of the CLEF 2011 tasks and also provides estimates on the applied human effort.

Of 18 tasks in the CLEF 2011 labs, seven exploited existing annotations in their collections to automatically generate relevance assessments, whereas 11 tasks employed human assessors. For the latter, four employed crowd sourcing for creating the human relevance assessments; three of them employed Amazon's Mechanical Turk, while one (Plant Identification at ImageCLEF) relied on another form of crowd sourcing, the members of a French social collaborative network in botany. The remaining seven tasks enlisted the help of 4-25 human assessors, mostly volunteers, e.g., students, task organizers, or even task participants, apart from the medical image retrieval task that recruited medical doctors and the music categorisation task that recruited music consultants, given the specialized nature of the domain of these tasks. Compared to 2010, the same number of tasks used crowdsourcing, but more tasks relied either on automatically generated relevance assessments or on human assessors. In this latter case, 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 several weeks for a single task.

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

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

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 eighth time in 2011. A new database was created for the use in ImageCLEF 2011 to allow for new challenges. The collection used a subset of 231,000 images from the PubMed Central database containing in total over one million images. This set of articles contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals. The subset chosen includes all journals of BioMed Central, as these allow redistribution of the data. A set of imaging oriented journals that also allow redistribution were taken in addition to this. See Appendix IV: CLEF 2011 Labs Test Collections for further details

Two main novelties of the new data set are that (1) there is a large variety of journals, not only radiology, meaning that rigor in figure legends is different from each other and the variety of images is much larger (ImageCLEF 2010 collection contained only 77,506 images) and that (2) the data set contains a majority of images that are not or little important for retrieval (such as tables, flow charts, graphs, etc.).

As in ImageCLEF 2010, three sub–tasks were conducted by the medical task: *medical modality classification*, *medical image retrieval* and *medical case retrieval.* In 2011, a new record of 130 research groups registered for the four sub-tasks of ImageCLEF down from seven sub tasks in 2009 but the same as in 2010. For the medical retrieval task the number of registrations also reached a new maximum with 55. 17 of the participants submitted results to the tasks, essentially the same number as in previous years. The following groups submitted at least one run:

- BUAA AUDR (BeiHang University, Beijing, China)
- CEB (National Library of Medicine, USA)
- DAEDALUS UPM (Universidad Politecnica de Madrid, Spain)
- DEMIR (Dokuz Eylul University, Turkey)
- HITEC (Ghent University, Belgium)
- IPL (Athens University of Economics and Business, Greece)
- IRIT (Institut de Recherche en Informatique Toulouse, France)
- LABERINTO (Universidad de Huelva, Spain)
- SFSU (San Francisco State University, USA)
- medGIFT (University of Applied Sciences Western Switzerland, Switzerland)
- MRIM (Laboratoire d'Informatique de Grenoble, France)
- Recod (Universidade Estadual de Campinas, Brazil)
- SINAI (University of Jaen, Spain)
- UESTC (University of Electronic Science and Technology, China)
- UNED (Universidad Nacional de Educacion a Distancia, Spain)
- UNT (University of North Texas, USA)
- XRCE (Xerox Research Centre Europe, France)

A total of 207 valid runs were submitted, 34 of which were submitted for modality detection, 130 for the image-based topics and 43 for the case-based topics. The number of runs per group was limited to ten per subtask and case-based and image-based topics were seen as separate subtasks in this view.

4.1 Medical Modality Classification Task

Previous research has demonstrated the utility of classifying images by modality in order to improve the precision of the search. In 2011, a simple ad-hoc hierarchy with 18 classes (see Table 2), 10 more classes than in 2010, was created for the Medical Modality Classification task. The sections radiology, microscopy, photography, graphics, other (see Figure 1) was created based on the existing data set.
For this hierarchy 1,000 training images and 1,000 test images were generated. Currently, a more detailed hierarchy based on a larger data set is being elaborated.

Table 2: Modality categories class codes with descriptions of the ImageCLEF 2011 medical modality classification task.

Class code	Description
AN	angiography
CT	Computed Tomography
MR	Magnetic Resonance imaging
US	Ultrasound
XR	X-Ray
FL	Fluorescence
EM	Electron Microscopy
GL	Gel
HX	Histopathology
PX	General photo
GR	Gross pathology
EN	Endoscopic imaging
RN	Retinography
DM	Dermatology
GX	Graphs
DR	Drawing
3D	3D reconstruction
CM	Compound figure (more than one type of image)

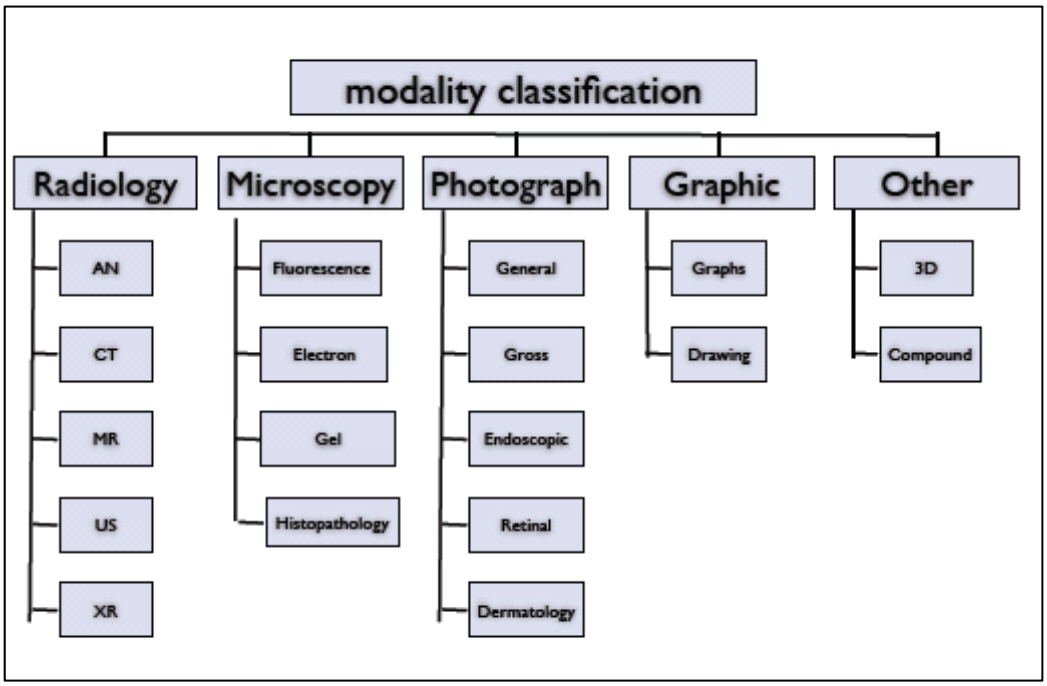

Figure 1: Modality categories of the ImageCLEF 2011 Medical Modality Classification task.

The results of the modality classification task are measured in classification accuracy. With a higher number of classes, this task was more complex than in 2010. Table 3 presents the top-10 results per run type (mixed, textual or visual). The best results were obtained by combining visual and textual methods (86%) as in 2010. The best run using visual methods (85%) had a slightly worse accuracy than the best run using mixed methods. The best run using textual methods alone obtained a much lower accuracy (70%). Only one single group submitted text-based results that performed worse than visual and mixed runs. Further details can be found in [1].

Table 3: Top-10 results per run type for the 2011 ImageCLEF Medical Modality Classification task.

Run Id	Group	Run Type	Classification Accuracy
CE_all_MIX_semiLM.txt	XRCE	Mixed	0.8691
XRCE_Testset_MIX_semiL50.txt	XRCE	Mixed	0.8642
2011.06.10-02.38.40.test.prediction.trec	HITEC	Mixed	0.8603
2011.06.09-18.36.25.test.prediction.trec	HITEC	Mixed	0.8564
XRCE_Testset_MIX_semiL25.txt	XRCE	Mixed	0.8593
2011.06.08-19.58.41.test.prediction.trec	HITEC	Mixed	0.8515
2011.06.10-00.01.26.test.prediction.trec	HITEC	Mixed	0.7685
Image_text_test_result_multilevel.dat	CEB	Mixed	0.7412
2011.06.10-03.25.40.test.prediction.trec	HITEC	Mixed	0.7412
Image_text_test_result_sum_ext.dat	CEB	Mixed	0.6025
ICLEF2011_MED_MODALITY_09062011_1500.txt	IPL	Textual	0.7041
ICLEF2011_MED_MODALITY_09062011_1600.txt	IPL	Textual	0.4765
XRCE_all_VIS_semiL25.txt	XRCE	Visual	0.8359
XRCE_Testset_VIS_semi20_CBIR.txt	XRCE	Visual	0.8349
XRCE_all_VIS_semi20_CBIR.txt	XRCE	Visual	0.8339
recod_imageclefmed_ModCla_357I	Recod	Visual	0.6972
recod_imageclefmed_ModCla_VI	Recod	Visual	0.6943
recod_imageclefmed_ModCla_VINoR	Recod	Visual	0.6904
recod_imageclefmed_ModCla_VsNoR	Recod	Visual	0.6835
recod_imageclefmed_ModCla_Vs	Recod	Visual	0.6806
recod_imageclefmed_ModCla_343s	Recod	Visual	0.6787
recod_imageclefmed_ModCla_370I	Recod	Visual	0.6787

4.2 Medical Image Retrieval Task

This is the classic medical retrieval task, similar to those in organized in 2005-2010. 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.

The topics for the image-based retrieval task were a selection of topics that had-been used in the past based on [2] [3]. Ten topics each for visual, textual and mixed retrieval were chosen to allow for the evaluation of a large variety of techniques. The reuse of existing topics allows for the comparison of the difficulty of these topics with various databases and limits the effort needed to survey clinicians and develop new topics. This also means that

participants had in principal various database available for training their systems, which can potentially increase performance.
Table 4, Table 5 and Table 6 present the top-10 results per run type (mixed, textual or visual). As in most previous years, the best results for the image-based retrieval topics were obtained using multimodal methods. Most of the runs submitted to this task use textual methods that perform well. 26 of the 130 submitted runs used purely visual techniques but the results were still much lower than the textual and multimodal techniques. This year, the multimodal run with the highest MAP (0.23) obtained better results than visual and textual techniques alone. In general the average performance of multimodal runs is lower than for purely textual retrieval underlining the importance of good fusion techniques. Further details can be found in [1].

Table 4: Top-10 results of the multimodal runs for the 2011 ImageCLEF Medical Image Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
mixed_3_2_cedd_baseline_run	DEMIR	Not applicable	0.2372	0.3933	0.2738
mixed_cedd_baseline_run	DEMIR	Not applicable	0.2307	0.3967	0.2606
mixed_3_2_cedd_weighted_run	DEMIR	Not applicable	0.2014	0.3400	0.2481
mixed_3_2_cedd_rerank_reindex_run	Mixed	Feedback	0.1983	0.4067	0.2428
mixed_cedd_weighted_run	DEMIR	Not applicable	0.1972	0.3367	0.2383
mixed_cedd_rerank_reindex_run	DEMIR	Not applicable	0.1853	0.3667	0.2230
DEMIR_MED2011	DEMIR	Automatic	0.1645	0.3967	0.2198
XRCE_RUN_MIX_SFLMODSc_ax_dir_spl	XRCE	Feedback	0.1643	0.3800	0.2234
XRCE_RUN_MIX_SFLMOD_ax_dir_spl	XRCE	Feedback	0.1545	0.3800	0.2053
XRCE_RUN_MIX_SFLMODFL2_ax_dir_spl	XRCE	Automatic	0.1520	0.3633	0.2049

Table 5: Top-10 results of the textual runs for the 2011 ImageCLEF Medical Image Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
laberinto_CTC	LABERINTO	Automatic	0.2172	0.3467	0.2402
Run2_Txt	UNED	Automatic	0.2158	0.3533	0.2514
IPL2011AdHocT1-C6-M0_2-R0_01-DEFAULT	UPL	Automatic	0.2145	0.4033	0.2434
laberinto_BC	LABERINTO	Automatic	0.2133	0.3400	0.2384
IPL2011AdHocT1-C6-M0_2-DEFAULT	IPL	Automatic	0.2130	0.3567	0.2370
Run3_Txt	UNED	Automatic	0.2125	0.3867	0.2430
IPL2011AdHocT0_113-C0_335-M0_1-DEFAULT	IPL	Automatic	0.2016	0.3733	0.2269
IVSCT5G	MRIM	Automatic	0.2008	0.3033	0.2331
IVSCT5GK	MRIM	Automatic	0.2008	0.3033	0.2331
IVPCT5GKin	MRIM	Automatic	0.1975	0.2967	0.2257

Table 6: Top-10 results of the visual runs for the 2011 ImageCLEF Medical Image Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
IPL2011Visual-DECFc	IPL	Automatic	0.0338	0.1500	0.0717
IPL2011Visual-DEFC	IPL	Automatic	0.0322	0.1467	0.0715
IPL2011Visual-DEC	IPL	Automatic	0.0312	0.1433	0.0716
ILP2011Visual-DEF	IPL	Feedback	0.028	0.136	0.070

			3	7	3
gift_visual_ib	medGIFT	Automatic	0.0274	0.1467	0.0807
ILP2011Visual-DTG	IPL	Automatic	0.0253	0.1333	0.0715
visual_ib	medGIFT	Automatic	0.0252	0.1267	0.0752
iti-lucene-image	CEB	Automatic	0.0245	0.1333	0.0627
image_fusion_category_weight_filter	CEB	Automatic	0.0221	0.1167	0.0651
image_fusion_category_weight_filter_merge	CEB	Automatic	0.0201	0.1000	0.0629

4.3 Medical Case Retrieval Task

This task was first introduced in 2009. This is a more complex task, but one that we believe is closer to the clinical workflow. In this task, a case description, with patient demographics, limited symptoms and test results including imaging studies, was provided (but not the final diagnosis). The goal is to retrieve cases including images that might best suit the provided case description and could be of help in differential diagnosis. Unlike the ad-hoc task, the unit of retrieval here is a case, not an image.

Table 7, Table 8 and Table 9 present the top-10 results per run type (mixed, textual or visual). As in 2010, almost all teams used textual retrieval techniques in the case-based retrieval task. Only one group submitted visual case-based retrieval runs. Best results were obtained with a textual retrieval approach. Multimodal fusion runs do not perform as well as text retrieval runs. Further details can be found in [1].

Table 7: Results of the multimodal runs for the 2011 ImageCLEF Medical Case Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
mixed_GIFT_Lucene_fulltext_cb	medGIFT	Automatic	0.0754	0.1667	0.0958
iti-lucene-baseline+expanded-concepts+image	CEB	Automatic	0.0269	0.0333	0.0252

iti-lucene-baseline+expanded-concepts+image+cases	CEB	Automatic	0.0255	0.0333	0.0230
iti-lucene-expanded-concepts+image	CEB	Automatic	0.0247	0.0333	0.0249

Table 8: Top-10 results of the textual runs for the 2011 ImageCLEF Medical Case Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
UESTC_full_indri	UESTC	Automatic	0.1297	0.1588	0.1212
HES-SO-VS_CASE_BASED_FULLTEXT	MedGIFT	Automatic	0.1293	0.1509	0.1122
UESTC_full_p2QE	UESTC	Automatic	0.1199	0.1365	0.1082
UESTC_full_p2	UESTC	Automatic	0.1179	0.1490	0.1162
MRIM_KJ_A_VM_Sop_T4G	MRIM	Automatic	0.1114	0.1546	0.1064
IRIT_LGDc1.0_KLbfree_d_20_t_20_1	IRIT	Automatic	0.1030	0.1206	0.0930
IRIT_CombSUMc1.0_KLbfree_d_20_t_20_1	IRIT	Automatic	0.0947	0.1073	0.0862
iti-essie-manual	CEB	Manual	0.0941	0.1409	0.1162
IRIT_LGDc1.0_KLbfree_d_20_t_20_1_ignore_low_idf	IRIT	Automatic	0.0937	0.1017	0.0716
MRIM_KJ_A_VM_Pos_T4G	MRIM	Automatic	0.0911	0.1454	0.0938

Table 9: Results of the visual runs for the 2011 ImageCLEF Medical Case Retrieval task.

Run Id	Group	Run Type	MAP	P10	bPref
gift_visual	medGIFT	Automatic	0.0204	0.0444	0.0292
bovw_visual_cb	medGIFT	Automatic	0.0164	0.0556	0.0267
_visual_ib	medGIFT	Automatic	0.0150	0.0444	0.0228
bovw_s2_visual_cb	medGIFT	Automatic	0.0082	0.0333	0.0113

4.4 Summary of the outcomes of the "Visual Clinical Decision Support" Use Case

The main outcomes of the second year evaluation activities for the "Visual Clinical Decision Support" use case realised within the ImageCLEFmed task are:

1. One year more, ImageCLEFmed was the most popular lab (see Section 2.1.2 and Appendix II: Participation in the CLEF 2011 labs). The number of registrations reached a new maximum with 60, 9 more than in CLEF 2010. 17 teams submitted at least one run in 2011, slightly more than in 2010. The numbers of runs increased from 155 to 207. There were more submissions on image-based retrieval task (130) than in the other two tasks modality classification (34) and case-based retrieval (43).
2. Similar to 2010, combination of evidence from various modalities is the most effective approach for the modality detection and medical image retrieval tasks, whereas textual methods are most effective for the medical case retrieval task. In particular:
   i. The modality classification task, with a higher number of classes, was more complex than in 2010. As seen in Table 3, the best results were obtained by combining visual and textual methods (86%) as in 2010. The best run using visual methods (85%) had a slightly worse accuracy than the best run using mixed methods. The best run using textual methods alone obtained a much lower accuracy (70%).
   ii. As in most previous years, the best results for the image-based retrieval task were obtained using multimodal methods. Most of the runs submitted to this task use textual methods that perform well. 26 visual runs were submitted but the results were still much lower than the textual and multimodal techniques (see Table 4Table 5 andTable 6).
   iii. As in 2010, for the medical case retrieval task, textual methods were clearly superior although only one group submitted visual case-based retrieval runs. Visibly, further research is needed for this task.

5 Outcomes of the evaluation activities for the "Search for Innovation" Use Case

As previously mentioned, the "Search for Innovation" use case involves searching in patent collections for making state-of-the-art assessments on a technical subject, at a given point in time. The CLEF-IP lab – benchmarking retrieval in the intellectual property – puts the evaluation activities in this use case in the foreground.

In 2011, CLEF-IP organized 5 tasks, corresponding to parts of the patent examination process: *Prior Art Candidate Search* (PAC), *Patent Classification Task* (CLS1), *Refined Patent Classification Task* (CLS2), *Image-based Patent Retrieval* (IMG-PAC), and *Image-based Classification* (IMG-CLS). The image-related tasks were organized in collaboration with the organizers of the ImageCLEF lab. We detail the participation in and the results of each of these tasks in the following subsections. In the following table we present the participating groups and mark the tasks to which experiments were submitted.

Table 10: List of participants and task submission to CLEF-IP 2011.

ID	Institution	Country	PAC	CLS1	CLS2	IMG-PAC	IMG-CLS
chemnitz	Chemnitz University of Technology, Retrieval Group	DE	x				
hildesheim	Hildesheim Univ. – Information Science	DE	x				
hprussia	Hewlett-Packard Labs, Russia	RU	x				
hyderabad	International Institute of Information Technology – SIEL	IN	x				
joanneum	Joanneum Research, Institute for Information and Communication Technologies	AT					x
lugano	University of Lugano	CH	x				
nijgmenen	Radboud University Nijgmenen, Information Foraging Lab	NL	x	x	x		
spinque	Spinque	NL	x				
tuwien-1	Vienna University of Technology, Inst. For Computer-Aided Automation	AT					x
tuwien-2	Vienna University of Technology, Inst. For Software Technology and	AT	x				x

	Interactive Systems						
wisenut	WISEnut Inc.	KR	x	x	x		
xerox	Xerox Research Centre Europe	FR				x	x
	Total runs submitted		**30**	**16**	**9**	**10**	**12**

5.1 Prior Art Candidates Search Task

The objective of the prior art candidate search task has not changed significantly from the one defined in 2010 (see also [1]). We state it here for completeness: retrieve documents from a collection of patents that could constitute prior art for a given topic patent. In other words, search for documents with technical details similar to (parts of) the technical details described in the given patent. The collection of patents used in the CLEF-IP lab contains documents with content in at least one of the following languages: English, German or French.

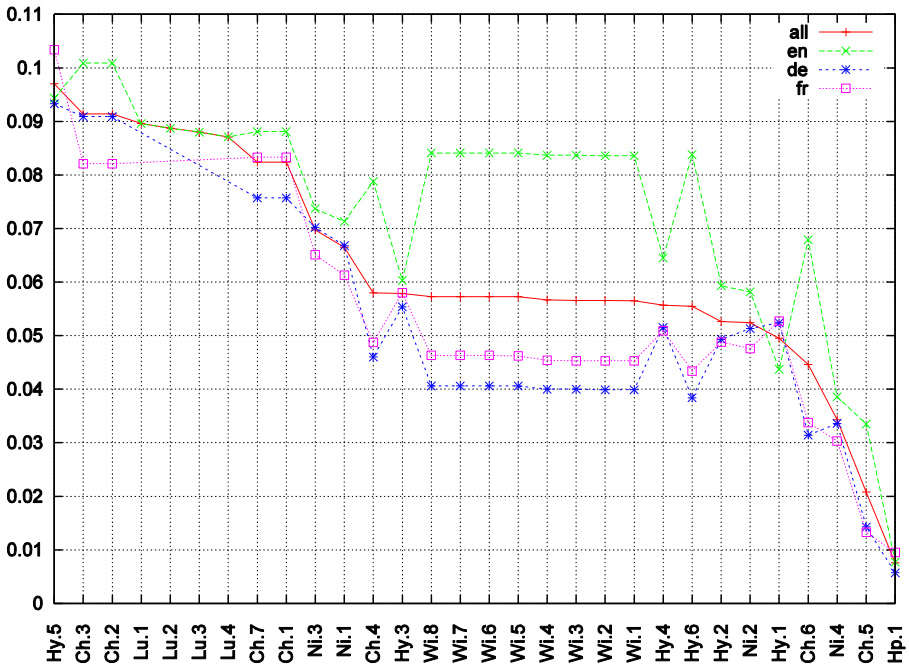

Figure 2: MAP scores for the 2011 CLEF-IP PAC task.

The topic set in 2011 consisted of 3973 topics, where differently from the task organized in 2010, the three collection languages, German, English and French, were equally represented: one third of the topics were in English, one third in German and one third in French. The relevance assessments for these topics were extracted from patent search reports produced by patent experts at patent offices. Such a search report lists the relevant patent documents found by a patent expert examining the original patent document.

Participants were allowed to use 2010 data as training data, in addition to a small set of 300 training topics made available in 2011. The 30 retrieval experiments submitted by the 9 task

participants were in textual format, with each line containing an answer to a given topic. At most 1000 answers per one topic were allowed.

Much of the work done by the participants was concentrated on how to generate queries out of the documents given as topics (involving various linguistic methods), and analysing how these impact the search results. Cross-lingual search reduced to translating queries into English, some participants treated only the English set of topics. Metadata in the documents (both collection and topics) was used to construct better queries, restrict the search space, or filter out retrieval results.

The measures computed for each of these runs are:

- precision at various cut-offs (5, 10, 20, 50, 100, all result set);
- recall at various cut-offs (5, 10, 20, 50, 100, all result set);
- Mean average precision (MAP);
- Normalized Discounted Cumulative Gain (NDCG).

Figure 2 shows the MAP scores for the whole set of topics, as well as for the three language-based sets of topics, for each of the received experiments in the PAC task. For further computations see [4].

5.2 Patent Classification Tasks

In 2011, CLEF-IP organized two patent classification tasks. The goal of these tasks was to classify given patent documents according to the International Patent Classification system (IPC[5]). The first of the two classification tasks, CLS1, was similar to the patent classification task organized in CLEF-IP 2010, which required classifying the documents at the IPC *subclass* level. The difference consists in that the three collection languages were equally distributed in the topic set, which contained 3000 topics. That is, a third of the topic documents were in English, a third in German, and a third in French.

The second classification task organized in CLEF-IP 2011, Refined Classification Task CLS2, asked the participants to classify a given patent document at the IPC *group/subgroup* level when the *subclass* classification was given.

To train their classification systems, participants were allowed to use only the documents in the CLEF-IP 2011 corpus. The relevance judgements for the topics were automatically extracted from the classification codes recorded in the original documents on which the topic documents were based.

To assess how the participating systems performed in these two classification tasks, we have computed the precision, recall and F_1 measures, each at cut-off levels 1 and 5. The obtained scores are shown in **Errore. L'origine riferimento non è stata trovata.** (CLS1 task) and **Errore. L'origine riferimento non è stata trovata.** (CLS2 task).

The two participants in the tasks have used various solutions to the classification problem, both with very good results. One participant used a linguistic classification system to implement three classifiers, which were further tuned to improve results. In representing the documents, it was experimented with various document data, also metadata, to include in

[5] The IPC system is a classification system organized hierarchically in sections, classes, subclasses, (main) groups and subgroups. It is maintained by the World Intellectual Property Organization.

the representation; parsed dependencies were also added to the representation. The second participant combined retrieval and classification algorithms in order to overcome the 'too little data to train on' issue – for underrepresented IPC classes, and 'not enough memory to contain the language model' issue – for the IPC classes with many documents. The first participant chose to treat only the English language topics; the second one has used the MyMemory[6] service to translate all non-English topics into English.

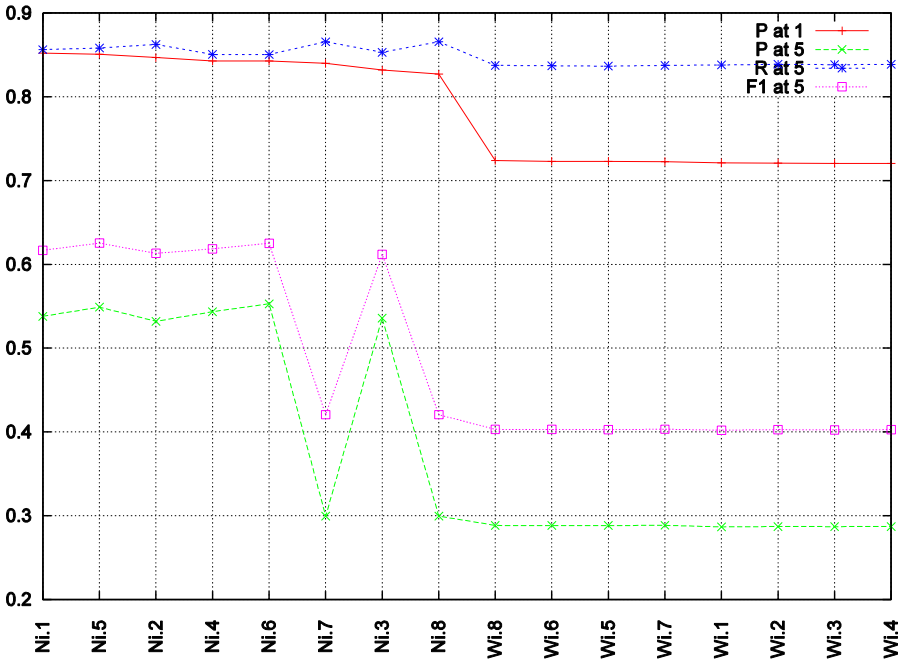

Figure 3: Precision, Recall and F1 scores for the 2011 CLEF-IP CLS1 task.

[6] http://mymemory.translated.net/

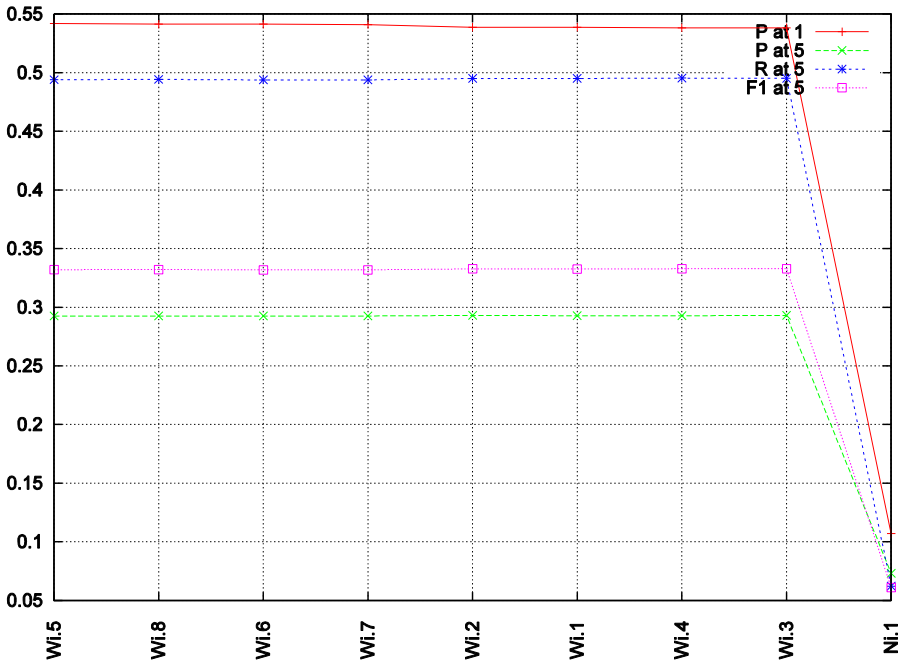

Figure 4: Recall and F1 scores for the CLEF-IP 2011 CLS2 task

5.3 Image-based Patent Retrieval

In many technological areas, patent professionals rely on images to either make a quick decision on the relevancy of a patent to some specific topic, or to clarify aspects of a patent under examination. For this reason, in 2011, CLEF-IP introduced a pilot task where patent retrieval involves patent images, in addition to the textual content of the documents (IMG-PAC).

The data collection for this task was restricted to three IPC subclasses (

Table 11), for which patent searchers very often rely on image comparisons to find relevant prior art. The restricted corpus included patent documents and the images attached to them. The principal reason for the restriction is that the task was organized as a pilot task, and that image data require very large storage space.

Participants to the task were asked to find, within the given, restricted corpus, the relevant documents for topics, where the topics consisted of text and images related to the text. There were 211 topics with usually more than one attached image to a topic textual document. The relevance judgements for this task were obtained in the same way as those for the PAC task.

The participation in this task was lower than expected, with only one group submitting retrieval experiments. The submission, however, supports our opinion that textual and image-based retrieval, combined, perform better than textual or image-based retrieval alone.

Table 11: IPC subclasses in the CLEF-IP 2011 corpus for the IMG-PAC task.

Subclass	Description
A43B	Characteristic features of footwear; parts of footwear
A61B	Diagnosis; surgery; identification
H01L	Semiconductor devices; electric solid state devices not otherwise provided for

5.4 Patent Image Classification

This was another new, image-based task in the CLEF-IP 2011. The aim was to automatically classify patent images based on visual content. Differently from all other tasks, the image-based patent classification task did not use patent documents and their textual content, but only black and white images extracted from patents. The participants were asked to classify the images into nine classes: drawing, chemical structure, program listing, gene sequence, flow chart, graph, mathematics, table, and symbol. For each of these classes training data were provided, with at least 300 and at most 6000 images per class. The test topic set contained 1000 images to classify.

To evaluate the performance of the classification systems we computed three measures: equal error rate (ERR), area under curve (AUC) of a ROC curve (Receiver Operating Characteristic curve), and true positive rate (TPR).

5.5 Other Activities Related to the 'Search for Innovation' Use Case

As in 2010, in 2011 steps were taken to collaborate with other campaigns involving patent retrieval, in particularly TREC-CHEM and PatOlympics. Two significant outcomes are a direct result of these efforts: First, the 2012 CLEF-IP has attracted domain experts who help us organise a chemical image segmentation and recognition task, thus tackling one of the problematic points of professional users working on chemical patents. Second, thanks to the addition of new expertise into the consortium (notably the groups at Sheffield and the Royal School of Library and Information Science in Copenhagen), the PatOlympics takes a new, more scientifically correct approach to observing users in practice.

5.6 Summary of the Outcomes in the 'Search for Innovation' Use Case

The main outcomes of the second year of evaluation activities for the 'Search for Innovation' use case are:

1. Taking a closer look at the results and methods obtained in the Prior Art Candidates Search Task we see that there is a need to at least reformulate this task such as to focus on specific tasks of a patent professional. Two important such tasks involve:
    a. searching iteratively with query refinement;
    b. claims having a final role in deciding the relevance degree of a candidate to prior art document.

Neither of these is currently reflected in the CLEF-IP Lab.

2. The very good scores the participants obtained in the classification tasks, both in 2010 and 2011, seem to show that patent classification, at least up to the subclass level of the IPC system, is an easy task. We will not organize a similar task in 2012. The same is valid for the patent image classification task as well, where the very good scores obtained by the two participants show that the existing research results in image processing (w.r.t. classification) can be successfully applied to patent images as well. We will not provide a patent image classification task in the next CLEF-IP Lab.
3. It is clear that patent images should be included into the search for prior art processes. It appears that the big majority of research groups, in order to obtain good results, work either on textual processing, or on image processing, but not on both. We have also recognized that the set up of the image-based patent retrieval task made it difficult to tackle. Breaking this task in finer-grained (sub)tasks involving image processing may be a better way to approach patent retrieval using patent images.
4. Involvement of patent professionals with this use case, be it as workshop participants, members of advisory board, or simply via personal contacts has substantially contributed to motivating CLEF-IP Lab and PatOlympics participants. At the same time it provided feedback to the work done in the use case. Concurrently, members of the EPO have recognized the relevancy of the research within the CLEF-IP Lab to more readily provide expanded support in 2012.

6 Outcomes of the evaluation activities for the "Unlocking Culture" Use Case

Documents in the cultural heritage (CH) domain are often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of various levels of complexity. CH institutions have various approaches to managing information and serve diverse user communities, often with specialized needs.

Recently developed and evaluated multilingual CH information systems (e.g. TEL[7], Multimatch[8], CACAO[9]) have focused on classical text retrieval of mostly bibliographic data. However, alternative retrieval and interaction scenarios such as browsing using timelines and geo-spatial searching, a strong focus on named entity searching, and exploratory searching are also of particular interest in the CH domain. Content providers and system producers in the CH domain might need alternative task and evaluation measures, for example the presentation of a variety of media within the search results or the provision of relevant contextual search functionalities (i.e. related items) for query reformulation or secondary searching.

In 2011, the CHiC2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage workshop[10] investigated evaluation efforts in the cultural heritage field as well as defining user scenarios and identifying possible relevant metrics for a benchmark CLEF lab. Representatives from the following institutions participated in the workshop:

- Humboldt Universität zu Berlin
- University of Amsterdam
- University of Padua
- University of Sheffield
- The Netherlands Institute for Sound and Vision
- Ionian University
- Swedish Institute of Computer Science, SICS

Participants were asked to bring in statements dealing with the following topics:

- use cases, evaluation needs, and best practices coming from field experience in the cultural heritage institutions;
- evaluation perspectives, frameworks, and approaches in the digital library and digital curation fields;

[7] http://search.theeuropeanlibrary.org/portal/en/index.html

[8] http://www.multimatch.eu/

[9] http://www.cacaoproject.eu/

[10] http://www.promise-noe.eu/chic-2011/home

- synergies and relationships between large-scale evaluation campaigns and CH evaluation.

Various talks addressed further challenges and possibilities of alternative evaluation activities.
The Digital Library Evaluation Ontology[11] (DiLEO) that integrates concepts of the digital library evaluation domain and their relations was introduced and discussed.

Discussions focussed on improved forms of interaction, how user-generated content can be leveraged and the impact on evaluation processes. Especially the entertainment component characteristic for CH environments requires other evaluation approaches and appropriate measures and metrics.

For 2012, the CHiC 2012[12] pilot evaluation lab aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems, creating evaluation tasks that represent the particular contingencies of the cultural heritage domain and should support system developers in defining systematic evaluation standards.

Experiences and results derived from CHiC2012 will inform next years' cultural heritage evaluation approaches within PROMISE and CLEF. In particular, the evaluation efforts are planned to be expanded to interactive and log-based approaches.

[11] http://dlib.ionio.gr/~gtsak/dileo/

[12] http://www.promise-noe.eu/chic-2012/home

7 Impact analysis for the CLEF initiative task

In the PROMISE Deliverable 6.1 [1] we introduced a preliminary scholarly impact analysis on the ImageCLEF campaign. After having obtained very interesting results we include a new task into the PROMISE WP6 to extend this work to a more detailed analysis of ImageCLEF and also an analysis of the impact of all of CLEF.

RSLIS (Royal School Of Library and Information Science in Copenhagen, Denmark) joined PROMISE in month 19 (spring 2012) and they are the responsible of the new task 6.6 "Impact analysis for the CLEF initiative". This task will focus on extending the initial analysis of the scholarly impact of evaluation campaigns carried out in the first half of the project. Specifically, this work will be extended by the following:

1. Developing a method for identifying source publications and citing publications more comprehensively and requiring less manual validation. This work will be performed in collaboration with Professor Erhard Rahm's team in the University of Leipzig that has developed a tool for performing online citation analysis of computer science research. The collaboration has been initiated and the tools developed at University of Leipzig are being extended for use on CLEF data, and tested on a sample dataset.
2. When citation analysis is used for research evaluation raw citation counts are rarely used to compare units of analysis as these may vary widely between research fields. Therefore a suitable baseline is established to normalise differences due to various citation behaviour and volume. Such a baseline should also be considered when the unit is evaluation campaigns, but may be challenging to establish as these are often multi-disciplinary in nature. Work will be initiated in July 2012 to investigate if such baselines can be meaningfully established for multi-disciplinary evaluation campaigns.

8 Outlook on future evaluation activities: CLEF 2012

This section provides an outlook on the upcoming evaluation activities in the third year of PROMISE by outlining the steps taken towards the organization of the CLEF 2012 conference and its current status and by listing the selected labs. This section provides only a brief summary of these activities; further details will be provided in PROMISE Deliverable 7.9 "Third PROMISE Annual Conference and Proceedings".

CLEF 2012 conference[13]: The CLEF 2012 Conference on Multilingual and Multimodal Information Access Evaluation will take place in Rome, Italy, on September 17-20, 2012. As in 2011, this event is organized by PROMISE. CLEF 2012 is built on the format first introduced in 2010, CLEF 2012 will consist of an independent peer-reviewed conference on a broad range of topics in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test various aspects of mono and cross-language Information retrieval systems. Together, the conference and the lab series will maintain and expand upon the CLEF tradition of community-based evaluation and discussion on evaluation issues.In summary:

1. A total of 38 papers were submitted to CLEF 2012, almost double of the number of submissions for the previous conference.
2. 285 research groups were initially registered to CLEF 2012. There is an increase compared to both CLEF 2010 and 2011.
3. As in 2011, there will be two keynote talks*:* Tobias Schrec*k,* University of Konstan*z,* an*d* Peter Clar*k,* Vulcan Inc. will be presenting*.*
4. The conference proceedings will be published in the Springer Lectures Notes in Computer Sciences (LNCS) as in previous years.
5. The format of mixing the scientific sessions and the labs over the duration of 3.5 days will also be kept.

CLEF 2012 labs: Following the tradition of past CLEF campaign, lab proposals were accepted for two types of labs:

1. *Evaluation labs* that follow a "campaign-style" evaluation practice for specific information access
2. *Lab workshops* organized as discussion sessions to explore issues of evaluation methodology, metrics, and processes in information access and closely related fields.

The lab sessions at the conference will contain ample time for general discussion and engagement by all participants - not just those presenting campaign results and papers. Organisers should plan time for panels, demos, etc. where applicable.

After the selection process, eight labs were accepted at CLEF 2012 and three labs were rejected. Seven labs will follow the evaluation lab format and one lab will be run as a

[13] http://www.clef2012.org/

workshop. The CLEF 2012 labs are the following:

1. *CHiC Cultural Heritage in CLEF*[14]: is a pilot evaluation lab aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. Data test collections and queries will come from the cultural heritage domain (in 2012 data from Europeana) and tasks will contain a mix of conventional system-oriented evaluation scenarios (e.g. ad-hoc retrieval and semantic enrichment) for comparison with other domains and a uniquely customized scenario for the CH domain, i.e., a variability task to present a particular good overview over the various object types and categories in the collection targeted towards a casual user. Three task are organized:
   i. *Ad-hoc Retrieval Task:* to measure information retrieval effectiveness with respect to user input in the form of queries.
   ii. *Variability Task:* to return diverse objects and resemble the diversity tasks of the Interactive TREC track or the CLEF Image photo tracks.
   iii. *Semantic Enrichment Task:* retrieval for a query to semantically enrich the query and/or guess the user's information need or original query intent.
2. *CLEF-IP Information Retrieval in the Intellectual Property Domain*[15]: provides a large collection of XML documents representing patents and patent images. On this collection the following four tasks are organized:
   i. *Chemical Structure Recognition Task*: starting from TIFF images containing patent scans, to identify the location of the chemical structures depicted on these pages and, for each of them, return the corresponding structure in a chemical structure file format.
   ii. *Flowchart Recognition Task*: extraction of the information in flowchart images in a predefined textual format.
   iii. *Passage retrieval starting from claims:* starting from a given claim, retrieval of relevant documents in the collection and mark out the relevant passages in these documents.
   iv. *Matching claim to description in a single document (Pilot):* starting from the claims of a patent application, indication of the paragraphs in the application's description section that best explain the contents of the given claim.

[14] http://www.promise-noe.eu/chic-2012/home/

[15] http://ifs.tuwien.ac.at/~clef-ip/

3. *ImageCLEF Cross Language Image Retrieval Track*[16]: evaluates the cross-language annotation and retrieval of images by focusing on the combination of textual and visual evidence. Four challenging tasks are foreseen:
   i. *Medical task:* image modality classification and image retrieval with visual, semantic and mixed topics in several languages, using a data collection from the biomedical literature.
   ii. *Photo annotation and retrieval:* semantic concept detection and concept-based retrieval using Flickr data, and large-scale annotation using general Web data.
   iii. *Plant identification:* visual classification of leaf images for the identification of plant species.
   iv. *Robot vision*: semantic localisation of a mobile robot using multimodal place classification, with special focus on generalization.

4. *INEX Initiative for the Evaluation of XML Retrieval*[17]: has been pioneering structured retrieval since 2002, and will join forces with CLEF running five tasks:
   i. *Social Book Search:* study of the value of user-generated descriptions in addition to formal metadata on a collection of Amazon Books and LibraryThing.com data.
   ii. *Data Centric:* study of ad-hoc search and facetted search on a collection of Linked Data (DBpedia) tied to a large corpus
   iii. *(Wikipedia) Snippet Retrieval:* study of the generation of informative snippets with sufficient information to determine the relevancy of search results.
   iv. *Show Me Your Code:* participants will submit system components (in particular feedback) rather than results.
   v. *Tweet Contextualization:* retrieve of synthetic contextual information from Wikipedia in response to a tweet with a URL on a small terminal like a phone.
5. *PAN Uncovering Plagiarism , Authorship and Social Software Misuse*[18]: offers three tasks:
   i. *Plagiarism Detection*: features a new plagiarism corpus based on the ClueWeb09, the new search engine ChatNoir which indexes the corpus, the cloud-based algorithm evaluation architecture TIRA and for the first time, real plagiarism cases.

[16] http://www.imageclef.org/

[17] http://inex.mmci.uni-saarland.de/

[18] http://celct.fbk.eu/QA4MRE/

ii. *Author Identification:* identification of sexual predators in chat logs and on authorship verification. Moreover, it features for the first time real cases of disputed authorship.

iii. *Quality Flaw Prediction in Wikipedia:* newly introduced, and it is about identification of Wikipedia articles that contain certain information quality flaws. It generalizes the vandalism detection task of CLEF2011.

6. *QA4MRE Question Answering for Machine Reading*[19]: evaluates Machine Reading abilities through Question Answering and Reading Comprehension Tests. This lab offers three tasks:

   i. *QA4MRE*: reading of single documents and identification of the answers to a set of questions about information that is stated or implied in the text.

   ii. *Processing Modality and Negation for Machine Reading (Pilot)*: evaluation whether systems are able to understand extra-propositional aspects of meaning like modality and negation.

   iii. *Machine Reading of Biomedical Texts about Alzheimer (Pilot)*: setting questions in the biomedical domain with a special focus on the Alzheimer disease.

7. ***RebLab Online Reputation Management***[20]**:** deals with the image that online media project about individuals and organizations. The aim is to bring together the Information Access research community with representatives from the Online Reputation Management industry, with the goals of (i) establishing a five-year roadmap that includes a description of the language technologies required in terms of resources, algorithms, and applications; (ii) specifying suitable evaluation methodologies and metrics; and (iii) developing of test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial systems. Two shared tasks on Twitter data are offered:
   i. *Monitoring task*: thematical clusterization of tweets including a company's name as a step towards early alerting on issues that may damage the company's reputation.
   ii. *Profiling task*: annotation of tweets according to their polarity for reputation (i.e. as to whether their content has positive/negative implications for the company's reputation).
8. *CLEFeHealth Cross-Language Methods, Applications, and Resources for eHealth Document Analysis*[21]: is a one-day workshop on cross-language evaluation of methods, applications, and resources for eHealth document analysis with a focus on written and spoken NLP.

[19] http://celct.fbk.eu/QA4MRE/

[20] http://www.limosine-project.eu/events/replab2012/

[21] www.nicta.com.au/clefehealth2012/

9 References

[1] T. Tsikrika, H. Müller, P. Forner, M. Frieseke, F. Piori, M. Agosti, E. Di Buccio and R. Berendsen, *PROMISE Deliverable 6.1: Report on the outcomes of the first year evaluation activities,* 2011.

[2] P. Forner, *PROMISE Deliverable 7.5: Second PROMISE Annual Conference and Proceedings,* 2011.

[3] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[4] J. Kalpathy–Cramer, H. Müller, S. Bedricks, I. Eggel, A. G. Seco de Herrera and T. Tsikrika, “Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks,” in *CLEF*, Amsterdam, 2011.

[5] W. Hersh, H. Müller and J. Kalpathy-Cramer, "The ImageCLEFmed Medical Image Retrieval Task Test Collection," *Digital Imaging,* vol. 22, no. 6, pp. 648-645, 2009.

[6] H. Müller, C. Despont-Gros, W. Hersh, J. Jensen, C. Lovis and A. Geissbuhler, "Medical image analysis and retrieval, User testing and task analysis," in *Proceedings of the Medical Informatics Europe Conference (MIE 2006)*, Maastricht, 2006.

[7] F. Piori, M. Lupu, A. Hanbury and V. Zenz, “Retrieval in the Intellectual property Domain,” in *CLEF*, Amsterdam, 2011.

[8] G. M. Di Nunzio, J. Leveling and T. Mandl, "LogCLEF 2011 Multilingual Log File Analysis:Language identification, query classification, and success of a query," in *CLEF*, Amsterdam, 2011.

[9] M. Gäde, N. Ferro and M. Lestari Paramita, "CHiC 2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage," in *CLEF*, Amsterdam, 2011.

[10] H. Goëau, P. J. Bonnet, N. Boujemaa, D. Barthelemy, J.-F. Molino, P. Birnbaum, E. Mouysset and M. Picard, "The CLEF 2011 Plant Images Classification Task," in *CLEF*, Amsterdam, 2011.

[11] R. Morante and W. Daelemans, “Overview of the QA4MRE Pilot Task: Annotating Modality and Negation for a Machine Reading Evaluation,” in *CLEF*, Amnasterdam, 2011.

[12] S. Nowak, K. Nagel and J. Liebetrau, “The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks, Notebook Papers,” in *CLEF*, Amsterdam, 2011.

[13] N. Orio and D. Rizo, “Overview of MusiCLEF 2011,” in *CLEF*, Amsterdam, 2011.

[14] A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, R. Sutcliffe, C. Forascu and C. Sporleder, “Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation,” in *CLEF*, Amsterdam, 2011.

[15] V. Petras and P. Clogh, “Introduction to the CLEF 2011 Labs,” in *CLEF*, Amsterdam, 2011.

[16] M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein and P. Rosso, “Overview of the 3rd International Competition on Plagiarism Detection,” in *CLEF*, Amsterdam, 2011.

[17] M. Potthast and T. Holfeld, “Overview of the 2nd International Competition on Wikipedia Vandalism Detection,” in *CLEF*, Amsterdam, 2011.

[18] T. Tsikrika, A. Popecu and J. Kludas, “Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011,” in *CLEF*, Amsterdam, 2011.

[19] S. Argamon and P. Juola, "Overview of the International Authorship Identification Competition at PAN-2011," in *CLEF*, Amsterdam, 2011.

[20] V. Petras, P. Forner and P. D. Clough, "CLEF 2011 Labs ans Workshop," in *CLEF*, Amsterdam, 2011.

Appendix I: Questionnaires sent to CLEF 2011 Labs organizers

- CLEF 2011 Labs
  https://docs.google.com/spreadsheet/ccc?key=0Ag8HzgTbd9JRdE5FcHE5aVo0bHNEVEs5ZzlKM2h0N1E&pli=1#gid=0

- CLEF 2011 Labs: collections
  https://docs.google.com/spreadsheet/ccc?key=0Ag8HzgTbd9JRdFpZSVVKYmNGd1NTRHhuMUkzaEluRlE#gid=0

Appendix II: Participation in the CLEF 2011 labs

Table 12: Participation to the CLEF 2011 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	2	17	2	1	8	16	PROMISE evaluation infrastructure
CLEF-IP	Patent Image-based Classification	1	5	2	Not applicable	Unrestricted	12	PROMISE evaluation infrastructure
CLEF-IP	Patent Image-based Prior Art Search	1	10	1	Not applicable	8	10	PROMISE evaluation infrastructure
CLEF-IP	Prior Art Candidates Search	3	17	9	3	8	30	PROMISE evaluation infrastructure
CLEF-IP	Refined Patent Classification	1	17	2	Not applicable	8	9	PROMISE evaluation infrastructure

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
ImageCLEF	Medical Image Classification and Retrieval	8	60	17	5	10	207	ImageCLEF submission system
ImageCLEF	Photo Annotation	6[22]	48	18	8	5	79	ImageCLEF submission system
	Concept-based Photo Retrieval	1		4	Not applicable	10	31	ImageCLEF submission system
ImageCLEF	Plant Identification	1	45	8	Not applicable	4	21	ImageCLEF submission system
ImageCLEF	Wikipedia Image Retrieval	4	45	11	9	20	110	ImageCLEF submission system
LogCLEF	Multilingual	3	17	4	2	Not	Not	Not

[22] 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.

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Log File Analysis					applicable	applicable	applicable
MusiCLEF	Music Categorisation	1	20	0	Not applicable	Unrestricted	0	e-mail
MusiCLEF	Music Identification	1	20	2	Not applicable	Unrestricted	0	e-mail
PAN	Authorship Identification	1	31	13	Not applicable	Unrestricted	92	Rapidshare
PAN	Plagiarism Detection	2	30	11	6	Unrestricted	105	Rapidshare
PAN	Wikipedia Vandalism Detection	2	18	3	0	Unrestricted	3	Rapidshare
QA4MRE	Annotating Modality and Negation for a Machine Reading	1 of QA4MRE 9 of QA@CLEF	0	0	Not applicable	1	0	CELCT submission System

Lab	Task(s)	Number of years the task is part of CLEF	Registrations	Participations	Return participations	Submissions allowed per participant	Total submissions	Submission system
	Evaluation							
QA4MRE	Question Answering for Machine Reading Evaluation	1 of QA4MRE 9 of QA@CLEF	25	12	12	10	62	CELCT submission System

Appendix III: Main outcomes of the CLEF 2011 Labs

Table 13: Main advancements in the CLEF 2011 Labs

Lab	Task(s)	Task type	Main differences/advancements from 2010
CLEF-IP	Patent Classification	Classification	More topics Better language distribution among the topics.
CLEF-IP	Patent Image-based Classification	Classification	Not applicable
CLEF-IP	Patent Image-based Prior Art Search	Retrieval	Not applicable
CLEF-IP	Prior Art Candidates Search	Retrieval	Topics are not built-up documents, but actual patent applications.
CLEF-IP	Refined Patent Classification	Classification	Not applicable
ImageCLEF	Medical Image Classification and Retrieval	Retrieval and Classification	Larger, totally difference dataset
ImageCLEF	Photo Annotation	Classification	Sentiment concepts were added.
	Concept-based Photo Retrieval	Retrieval	Novel retrieval task.

Lab	Task(s)	Task type	Main differences/advancements from 2010
ImageCLEF	Plant Identification	Classification	Not applicable
ImageCLEF	Wikipedia Image Retrieval	Retrieval	More visual examples were provided in the topics. The average number of image examples per topic increased from 1.7 to 4.8. More topics with named entities and more specific topics were provided, since these two types of topics are representative or real web image search. The visual features extracted from the images in the collection and the image examples which were provided to the participants so as to support those coming from the textual IR community were improved compared to 2010. Crowdsourcing was applied for performing the relevance assessments.
LogCLEF	Multilingual Log File Analysis	Log Analysis	Generation of ground truth and sharing of resources
MusiCLEF	Music Categorisation	Classification	Not applicable
MusiCLEF	Music Identification	Retrieval	Not applicable
PAN	Authorship Identification	Classification	Not applicable
PAN	Plagiarism Detection	Retrieval	More difficult corpus More manually crafted plagiarism (crowdsourced) First time introduction of manually crafted translation plagiarism (crowdsourced).
PAN	Wikipedia Vandalism Detection	Classification	More languages were used.

Lab	Task(s)	Task type	Main differences/advancements from 2010
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation	Question Answering	Not applicable
QA4MRE	Question Answering for Machine Reading Evaluation	Question Answering	Major innovation New evaluation focus on the reading of a single document Use of background collections

Table 14: Main trends in the approaches employed by the participants to the CLEF 2011 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		
CLEF-IP	Patent Image-based Classification	Bags of keywords, Fisher Vectors, Simple features	Best run had a True Positive Rate of 0.91, so the problem is basically solved. It will not be run again in this form.
CLEF-IP	Patent Image-based Prior Art Search		
CLEF-IP	Prior Art Candidates Search	Apply linguistic methods processing the data	
CLEF-IP	Refined Patent Classification		
ImageCLEF	Medical Image Classification and Retrieval	Query expansion was often successful, mapping to MeSH terms and using the MeSH hierarchy	• Multimodal approaches are often best • Visual has good early precision • Fusion is hard to do

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
ImageCLEF	Photo Annotation	SIFT features and discriminative approaches	• Performance of textual runs is close to that of visual runs • Multimodal approaches outperform classification with single modality information • 17 concepts were detected best with a visual approach, 3 concepts were detected best by a textual approach, and the remaining 79 with a multimodal approach.
	Concept-based Photo Retrieval		• Manual runs work best, independent from the configuration (textual, visual, multi-modal) • Great variability of performance for various topics

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
ImageCLEF	Plant Identification	Numerous shape based approaches	• Shape based approaches are more effective. But one matching approach was more relevant on scans • Metadata like gps, taxonomical context where not successfully exploited • Free photographs of leaves are very difficult to identify with state of the art methods and without manual interactions (like segmentation for instance)
ImageCLEF	Wikipedia Image Retrieval	• More multimodal (and multilingual approaches) are being developed/ used • Many groups use external sources to enhance retrieval (e.g., Flickr, WordNet etc.) • Trend to use components provided by other participating groups so that each group can focus only on the specific aspects of research that are of interest to them.	• Multilingual approaches are more successful than monolingual ones • For 8 of the 9 that submitted both multimodal and monomedia approaches, their multimodal outperformed mono-media runs. This is probably due to increased number of visual examples, improved visual features, and more appropriate fusion techniques • There were also many (successful) query/document expansion submissions.
LogCLEF	Multilingual Log File Analysis	Measurable effects on the success of search query based on language correlations	Native language vs. interface language may influence how user interacts with application, interface language changes during a session may give hint about user search preferences.

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
MusiCLEF	Music Categorisation	Not applicable	Not applicable
MusiCLEF	Music Identification	Not applicable	Not applicable
PAN	Authorship Identification	Corpus quality	Task is practical using standard methods.
PAN	Plagiarism Detection	• Exhaustive comparison of suspicious documents to source documents • Similar document indexing pipeline • Dotplot-based plagiarism extraction • Intrinsic detection based on outlier detection.	• Remarkable performance improvements for intrinsic detection • Some improvements, however, might not hold up in practice due to corpus deficiencies • External detection posed a renewed challenge because of the more difficult corpus • Few participants managed to perform well on all measures.
PAN	Wikipedia Vandalism Detection	• Continuing 2010's trends the best performance was achieved with a combination of content-based and context-based features • First time application of language-dependent features • First-time application of a-posteriori features	Detection performance improved significantly as a result of the new kinds of features employed

Lab	Task	Main trends (among the participants' approaches)	Main experimental outcomes (based on the participants' results)
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation	Not applicable	Not applicable
QA4MRE	Question Answering for Machine Reading Evaluation	Question answering systems plus answer validation	The task is affordable and more realistic that in past editions

Appendix IV: CLEF 2011 Labs Test Collections

List of collections in the CLEF 2011 labs:

1. **CLEF-IP 2011 collection**: A collection extending the CLEF-IP 2010 collection. It contains patent documents from the EPO (European Patent Office) that have an application date previous to 2002. In addition, for Euro-PCT applications, the corresponding patent documents published by the WIPO (World Intellectual Property Organization) were added. The files contain bibliographic data in addition to descriptive text. The XML files are quite comprehensive, containing detailed information on inventors, assignees, priority dates etc as well as invention-title, classifications-ipcr, abstract.
2. **CLEF-IP image classification 2011**: A collection of 38,081 training images and 1,000 test images taken from patents and organized into 9 classes. The collection was extracted from the MAtrixware REsearch Collection (MAREC), from which the datasets for CLEF-IP 2009 and 2010 were built.
3. **CLEF-IP-IMG 2011 collection**: A collection based on CLEF-IP 2011 collection but restricted to 3 IPC subclasses (A43B, A61B, H01L). Tif image files contain patent images attached to patent documents.
4. **PubMed Central images**: A collection of medical images obtained from PubMed Central. The database distributed includes XML file with the image and its id, the captions of the images, the titles of the journal articles in which the image had appeared and the PubMed ID of the journal article.
5. **MIR Flickr images collections**: Two collections of images obtained from the MIR Flick. The Flickr photos are collected based on interestingness rating, including Flickr user tags and EXIF tags for most of the photos. The annotations are provided as plain txt files.
6. **Pl@ntLeaves**: A collection based on the Plant@Leaves dataset which focuses on 71 tree species from French Mediterranean area. It contains around 5,436 pictures subdivided into 3 various kinds of pictures: scans, scan-like photos and free natural photos.
7. **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.
8. **Deutscher Bildungsserver (DBS) logs:** A collection of logs which are server logs in standards format in which the searches and the results viewed can be observed. The "Deutscher Bildungsserver" is a quality controlled internet directory for educational resources.
9. **The European Library (TEL) logs**: A collection of search/action logs stored in a

relational table and containing various types of actions and choices of the user. Each record represents a user action. Three years and a half of log data were released.

10. **Sogou logs**: A collection of query logs containing queries to the Chinese Sogou search engine.
11. **Autotagging**: A collection of songs stored in MP3 format and web crawled pages for artists and tags.
12. **Fonoteca, RTI**: A collection of thousands of songs in MP3 format, JPG cover images and metadata describing the albums. A company for music broadcasting services (LaCosa s.r.l.) and a public music library (University of Alicante's Fonoteca) provide the data. The collection is mostly biased towards pop and rock genres, although about 10,000 files are recordings of classical music and will be used for one of the tasks. To completely overcome copyright issues, only lowlevel descriptors were distributed to participants.
13. **Pan-11 Authorship Corpus**: A collection based on Enron Email corpus consisting of real-world texts.
14. **PAN Plagiarism Corpus 2011 (PAN-PC-11)**: A collection consists of documents in which a large number of plagiarism cases have been inserted. The plagiarism varies mainly with respect to the parameters length and obfuscation type.
15. **PAN Wikipedia Vandalism Corpus 2011 (PAN-WVC-11)**: A collection based on PAN-WVC-10. The collection contains a random sample from Wikipedia edit logs in 3 languages.
16. **QA4MRE 2011**: A multilingual collection of reading comprehension tests of given documents. Each test consists of one single document (Test Document) with several questions and a set of choices per question.
17. **QA4MRE-modality**: A collection containing test documents specifically selected from QA4MRE 2011 in order to ensure the properties required for the pilot task "Annotating Modality and Negation for a Machine Reading Evaluation".

Table 15: Collections used in the tasks of the CLEF 2011 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	Patent Classification Prior Art Candidates Search Refined Patent Classification	CLEF-IP 2011	~2,900,000	100GB	EN, DE, FR	Yes	1	The 2011 CLEF-IP data collection is based on the 2010 data, extending it with some WIPO patent documents. Both collections are extracted from the MAREC data corpus.
CLEF-IP	Patent Image-based Classification	CLEF-IP image classification 2011	39,081	380MB	Not applicable	Yes	1	The collection was collected from the MAREC patent collection, from which the datasets for CLEF-IP 2009 and 2010 were extracted.
CLEF-IP	Patent Image-based Prior Art Search	CLEF-IP – IMG 2011	~2,500,000	~5.3GB	EN, DE, FR	Yes	1	For this task, the 2011 CLEF-IP collection was restricted to 3 IPC subclasses (A43B, A61B, H01L).

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
ImageCLEF	Medical Image Classification and Retrieval	PubMed Central	231,000 images	16 GB	Mainly EN	Yes	1	None
ImageCLEF	Photo Annotation: Annotation task	a subset of the MIR Flickr dataset	18,000 photos		Mainly EN	Yes	3	The whole annotation task collection has been used before, but with less visual concepts. Also the ground truth had not been previously provided for the test set.
	Photo Annotation: Concept-based Photo Retrieval task	a subset of the MIR Flickr dataset	200,000 photos	~3 GB	Mainly EN	Yes	1	None

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
ImageCLEF	Plant Identification	Pl@ntLeaves	5,436 images and associated textual metadata	332 MB	EN, FR, LA	No	1	None
ImageCLEF	Wikipedia Image Retrieval	ImageCLEF 2010 Wikipedia	237,434 images and associated user-supplied annotations	25GB	EN, FR, DE	Yes	2	The collection was created for and used in ImageCLEF Wikipedia image retrieval task in 2010
LogCLEF	Multilingual Log File Analysis	DBS logs		5GB (zipped)	(mostly) DE, (some) EN	Yes	2	The DBS logs were used in LogCLEF 2010.
LogCLEF	Multilingual Log File Analysis	Sogou logs		1.9GB (zipped)	Any, mostly Chinese.	Yes	1	None
LogCLEF	Multilingual Log File Analysis	TEL logs	3580000	2GB (zipped)	Any, usually European.	Yes	1	The TEL logs were used in LogCLEF 2009, 2010.

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
MusiCLEF	Music Categorisation	Autotagging	1,355	7Gb	Any	Yes	1	None
MusiCLEF	Music Identification	Fonoteca + RTI	40,000 LP (Fonoteca) + 320,000 songs (RTI)	30 GB + 1.5Tb	ES, DE, IT, EN, FR	Yes	1	None
PAN	Authorship Identification	Pan-11 authorship corpus	large set: 9,337 docs small set: 3,001docs	2.3 MB zipped	EN	Yes	1	None
PAN	Plagiarism Detection	PAN Plagiarism corpus 2011 (PAN-PC-11)	26,939	4.6 GB	EN,DE, ES	Yes	1	None

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
PAN	Wikipedia Vandalism Detection	PAN Wikipedia Vandalism corpus2011 (PAN-WVC-11)	29949	8.4 GB	EN,DE, ES	Yes	1	None
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation	QA4MRE-modality	12	148 KB	EN	Yes	1	None
QA4MRE	Question Answering for Machine Reading Evaluation	QA4MRE 2011	919756	9.1 GB	EN, DE, ES, IT, RO	Yes	1	None

Table 16: Topics used in the tasks of the CLEF 2011 Labs.

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
CLEF-IP	Patent Classification	Classification	A patent application document, A1 or A2, where the classification information was removed.	3,000 docs 639 classes[23]	EN, DE, FR
CLEF-IP	Patent Image-based Classification	Classification	An image occurring in a patent application document.	1,000 images 9 classes	Not applicable
CLEF-IP	Patent Image-based Prior Art Search	Retrieval	A patent application document, A1 or A2, where the citation information was removed, AND images occurring in the application documents.	211 docs	EN, DE, FR
CLEF-IP	Prior Art Candidates Search	Retrieval	A patent application document, A1 or A2, where the citation information was removed.	3,973 docs (+300 training)	EN, DE, FR
CLEF-IP	Refined Patent Classification	Classification	A patent application document, A1 or A2, where the classification information was removed.	3,000 docs 20.000 classes[24]	EN, DE, FR
ImageCLEF	Medical Image Classification and Retrieval	Classification	An image from the medical literature.	1,000 images 18 classes	EN, FR, DE
		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, FR, DE

[23] These are all possible classes; out of them, 491 actually occurred in the test data

[24] These are all possible classes; out of them, 7,267 actually occurred in the test data.

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
ImageCLEF	Photo Annotation	Classification	A Flickr image.	10,000 images 99 concepts	EN
	Concept-based Photo Retrieval	Retrieval	A topic is a multimedia query that consists of a textual part, a Boolean connection of concepts, and a visual part, one or several example images.	40	EN
ImageCLEF	Plant Identification	Classification	A leaf picture.	1,440 images 70 species	EN, FR, LA
ImageCLEF	Wikipedia Image Retrieval	Retrieval	A topic is a multimedia query that consists of a textual part, the query title, and a visual part, one or several example images.	50	EN, FR, DE
LogCLEF	Multilingual Log File Analysis	Log Analysis	Queries in logs can be seen as topics.	~ 1,000,000 TEL records	Any
MusiCLEF	Music Categorisation	Classification	A music piece (to be categorised for possible usage)	380 music pieces 94 classes	EN, IT, DE, FR, SV
MusiCLEF	Music Identification	Retrieval	A music piece (to find other files in the dataset that are or contain versions (covers) of the same music piece).	600 single music pieces + 19 LP music pieces	(mostly) IT, ES
PAN	Authorship Identification	Classification	Email texts for which to identify the author.	~100 docs 1 author (yes/no) 400 docs 26 authors 1500 docs	EN

Lab	Task(s)	Task type	What constitutes a topic for this task?	Topics	Languages
				72 authors	
PAN	Plagiarism Detection	Retrieval	The document to be analysed plagiarism.	13,469	EN, DE, ES
PAN	Wikipedia Vandalism Detection	Classification	The Wikipedia article being edited.	30,000	EN, DE, ES
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation	Question Answering	A natural language question with 5 alternative answers.	120	EN
	Question Answering for Machine Reading Evaluation				EN, DE, ES, IT, RO

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Table 17: 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	All documents in the collection	Automatic relevance assessments		
CLEF-IP	Patent Image-based Classification	All documents in the collection	5	Volunteers (organisers, participants, others)	30 hours
CLEF-IP	Patent Image-based Prior Art Search	All documents in the collection	Automatic relevance assessments		
CLEF-IP	Prior Art Candidates Search	All documents in the collection	Automatic relevance assessments		
CLEF-IP	Refined Patent Classification	All documents in the collection	Automatic relevance assessments		

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
ImageCLEF	Medical Image Classification and Retrieval	~30,000 (pooling: top 50)	~15	Medical doctors in a medical information program in Portland OR, USA.	~250 hours
ImageCLEF	Photo Annotation	18,000 photos	414	Amazon Mechanical Turk workers	On average: about 2 minutes for the assessment of 10 images with 9 sentiment concepts. The other concepts were assessed in past ImageCLEF cycles
	Concept-based Photo Retrieval	56,909 photos (pooling: top 100)	hundreds	Amazon Mechanical Turk workers	On average: 31 seconds to 1 minute and 19 seconds for one HIT (consisting of 24 photos for one topic)

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
ImageCLEF	Plant Identification	5436	17	Members from Tela Botanica, the French social collaborative network in botany.	Each assessor collected leaves on trees, scanned and entered annotations through an online collaborative database system. Some of the assessors certainly spent hours and hours if we consider the collection time. An optimistic estimation: about 3 minutes per picture.
ImageCLEF	Wikipedia Image Retrieval	73,346 images (pooling: top 100)	379	Amazon Mechanical Turk workers	On average: about 27 minutes per topic
LogCLEF	Multilingual Log File Analysis	1,290 records manually annotated 940,957 records automatically annotated	25	LogCLEF participants	Not available (requested 50 annotations per assessor)
MusiCLEF	Music Categorisation	1355	Not available	LaCosa s.l.r	Several weeks
MusiCLEF	Music Identification	1355 songs + 22 LP	4	Volunteers (organisers, participants, others)	50 hours
PAN	Authorship Identification	All documents in the collection	Automatic relevance assessments		None; but ~300 hours generating tests in suitable format
PAN	Plagiarism Detection	All documents in the collection	Automatic relevance assessments		

Lab	Task(s)	How many documents were assessed?	How many assessors were employed?	Who were the assessors?	How much time did the assessors spend?
			(A couple thousand paraphrases were obtained via crowdsourcing)		
PAN	Wikipedia Vandalism Detection	All documents in the collection	Hundreds	Mechanical Turk workers	On average: 1-2 seconds per edit
QA4MRE	Annotating Modality and Negation for a Machine Reading Evaluation	All documents in the collection (the 120 questions)	5	Volunteers (organisers, participants, others)	2 months
	Question Answering for Machine Reading Evaluation				